Comparative analyses of adaptive phenomena in eukaryotic systems

Bush, Stephen

Award date:
2015

Awarding institution:
University of Bath

Comparative analyses of adaptive phenomena in eukaryotic systems

Stephen James Bush

A thesis submitted for the degree of Doctor of Philosophy

University of Bath
Department of Biology and Biochemistry

August 2014

I dedicate this thesis, with love and gratitude,

to my parents

Dawn and Ali

Acknowledgements

I'd like to thank my supervisor Araxi Urrutia in three-fold fashion, and in order: for the tremendous opportunity offered in joining her lab, for her guidance and patience in shaping this thesis, and finally, but foremost in my mind, for the faith placed in an unknown's ability. I'm privileged to have had this position and not only grateful for all that I've learned but knowing now what to do next. I'd like to thank also Bath's bioinformatics coterie (past and present) for reasons technical (of course) but rather more importantly, socially too (Claudia, Katie, Jaime, Lu, Nina, Marina, Atahualpa, Jimena, Wei...): it's been a stimulating environment, quite independent of my coffee excesses, and I couldn't have asked for it better! To Bath's resident tutor team, too, for all that was packed into, and around, the 7pm hour, and for the discreet little niche we all carved out on campus.

Following naturally from all the above, then, I take great pleasure in saying:

To my friends old and new for every added good memory, of which I'm glad to have plenty, and for everything that isn't particular – for the wine, the wit, the wandering and all sorts of coffees... for your companionship and assistance along this journey, in which I'd otherwise have disregarded the sun to favour computers, in all ways I'm honoured you were there.

To Ed, for our on-stage adventures, or 'PhD procrastination that got out of hand' and for all who supported and encouraged our projects, or otherwise committed directly: I'm delighted and proud in equal measure that together we've managed to achieve this!

To Elle, for a grounding on bedrock and all that's thereafter...

And finally, although with the greatest of gratitude, I'd like to thank my family – my parents Dawn and Ali, and my brother Kevin – for their love and support, both unwavering, in seeing this PhD through.

My love and thanks to all for joining this journey. I couldn't have achieved this without you.

Contributions

Results presented in chapters 3 and 4 have been published in the journal of Molecular Biology and Evolution:

Chapter 3 may be cited as: Bush, S.J., *et al.* (2014) Presence/absence variation in *A. thaliana* is primarily associated with genomic signatures consistent with relaxed selective constraints. Molecular Biology and Evolution. 31(1): 59-69.

Chapter 4 may be cited as: Chen, *et al.* (2014). Correcting for differential transcript coverage reveals a strong relationship between alternative splicing and organism complexity. Molecular Biology and Evolution. 31(6): 1402-1413.

Results presented in chapter 2 have been prepared and submitted for publication.

I confirm that I have designed and carried out all of the work presented in this thesis as well as, with the support of my supervisor Araxi Urrutia, the writing of all chapters presented, with the following exceptions:

Lu Chen and Jaime M. Tovar-Corona generated the alternative splicing annotations and calculated a comparable alternative splicing index, used in chapters 2, 3 and 4.

Lu Chen also contributed to the original design of the study presented in chapter 4; we are both co-first authors of the resulting publication.

Atahualpa Castillo-Morales contributed the functional category enrichment analysis included in chapter 3 and the phylogenetic generalised least squares regression included in chapter 4.

Paula Kover provided genomic data for multiple *A. thaliana* accessions, and contributed to the conception and design of chapters 2 and 3.

All of the aforementioned also contributed valuable edits and critical commentary to the wording of the manuscripts here presented.

Abstract

All phenotypic adaptations are encoded in the genome, although untangling the relationships between specific phenotypes and genomic changes is complicated by the fact that at the molecular level, most changes are associated not with selection but neutral processes. Distinguishing between the two is necessary for reliable interpretation of any potential signature of selection. In this respect, this thesis addresses three aspects of genome evolution: biasing factors for the interpretation of protein evolutionary rates, the adaptive significance of presence/absence variation (PAV) in the model plant *A. thaliana* and the evolution of alternative splicing alongside increasing organism complexity. A comparative genomics approach is used throughout, taking advantage of the increasing availability of high-throughput sequencing data. We show that *(i)* lineage-specific substitutions and the differential conservation of the edges of exons influence interpretations of protein evolutionary rate, *(ii)* that PAV in *A. thaliana* can be explained without invoking adaptation, despite enrichments for PAV events in genes considered as positively selected, and *(iii)* that alternative splicing is amongst the strongest predictors of organism complexity, consistent with an adaptive role of transcript diversification in determining a genome's functional information capacity. Taken together, we find that the signatures and targets of adaptation can either be masked by, or re-interpreted in the context of, non-adaptive processes and that with the increasing availability of high-throughput data, such considerations are of increasing relevance.

Contents

Acknowledgements ii

Contributions iii

Abstract iv

List of Figures vii

List of Tables ix

List of Supplementary Tables xi

Abbreviations xiv

1 Introduction 1

2 Lineage-specific sequence evolution and exon edge conservation partially explain the relationship of evolutionary rate and expression level in *A. thaliana* 9

2.1 Summary . . . 9

2.2 Introduction . . . 10

2.3 Results . . . 11

2.3.1 Correlates of dN/dS in *A. thaliana* . . . 11

2.3.2 Accounting for exon edge conservation influences dN/dS and its relationship with various genomic parameters, and unmasks higher levels of positive selection . . . 13

2.3.3 Using the more distant relative *T. parvula* results in similar patterns to those found with comparisons to *A. lyrata* . . . 15

2.3.4 Reduced prominence of gene expression as a predictor of *A. thaliana*'s lineage-specific dN/dS . . . 16

2.4 Discussion . . . 17

2.4.1 Lineage-specific substitution estimates and the conservation of exon edges partially explain the association between gene expression and dN/dS in *A. thaliana* . . . 17

2.4.2 Gene length is significantly associated with dN/dS values obtained from pairwise and lineage-specific substitutions for *A. thaliana* 19
2.4.3 Exon edge removal, but not lineage-specific substitution patterns, unmask higher levels of positive selection 20
2.5 Materials and Methods 22
2.5.1 Data sources 22
2.5.2 Tests of sequence evolution and selection 22
2.5.3 Exon edge trimming 23
2.5.4 Alternative splicing 23
2.5.5 Randomisation test 23
2.5.6 Expression data 24
2.5.7 Other data sources 24

3 Presence/absence variation in *A. thaliana* is primarily associated with genomic signatures consistent with relaxed selective constraints 26
3.1 Summary 26
3.2 Introduction 27
3.3 Results 28
3.3.1 Genes involved in signal transduction and both nucleotide and protein binding are over-represented among PAV genes 29
3.3.2 Genes affected by PAV show signatures consistent with relaxed selective constraints 35
3.3.3 PAV genes are located in genomic regions that are gene-poor and transposable element-rich 40
3.3.4 Exon loss is associated with a marginal reduction in expression level 41
3.4 Discussion 43
3.5 Materials and Methods 47
3.5.1 Genome sequence and annotations 47
3.5.2 Detecting missing exons relative to Col-0 48
3.5.3 Functional category enrichment analysis 48
3.5.4 Sequence evolution analysis 48
3.5.5 Gene expression 49
3.5.6 Paralogue number and gene age annotations 49
3.5.7 Transposable element and hotspot motif density 50
3.5.8 Alternative splicing events 50
3.5.9 Randomisation test 51

4 Correcting for differential transcript coverage reveals a strong relationship between alternative splicing and organism complexity **52**

4.1 Summary 52

4.2 Introduction 53

4.3 Results 55

4.3.1 Alternative splicing prevalence has increased throughout evolutionary time 55

4.3.2 Alternative splicing is a strong predictor of organism complexity, assayed as cell type diversity 58

4.4 Discussion 67

4.5 Materials and Methods 72

4.5.1 Organism complexity 72

4.5.2 Identification of alternative splicing events 72

4.5.3 Additional functional genomic parameters 72

4.5.4 Statistical analysis 73

4.5.5 Correction for phylogenetic autocorrelation 73

4.5.6 Expression level 74

5 General discussion **75**

Bibliography **83**

List of Figures

2.1 dN, dS, dN/dS and NI after exon edge removal. 14

3.1 Location within a Col-0 gene of exons missing in at least one other accession. 28

3.2 Distribution of 'exon presence/absence' (E-PAV) genes ($n = 330$) – those with at least one, but not all, exons missing in at least one accession – by GOslim categories for molecular function (a), biological process (b) and cellular component (c), and by family (d). 30

3.3 Distribution of 'CDS presence/absence' (CDS-PAV) genes ($n = 81$) – those with their entire coding region missing in at least one accession – by GOslim categories for molecular function (a), biological process (b) and cellular component (c). 31

3.4 Distribution of 'exon presence/absence' (E-PAV) genes – those with at least one, but not all, exons missing in at least one accession – by GO category ($n = 330$). 32

3.5 Distribution of 'exon presence/absence' (E-PAV) genes – those with at least one, but not all, exons missing in at least one accession – by Pfam category ($n = 330$). 34

3.6 Distribution of 'CDS presence/absence' (CDS-PAV) genes – those with their entire coding region missing in at least one accession – by Pfam category ($n = 81$). 35

3.7 Genetic features associated with intact (having no exons under P/A variation), E-PAV (having at least one, but not all, exons missing in at least one accession), and CDS-PAV (having the entire CDS missing in at least one accession) genes. 36

3.8 Distribution of Tajima's D values for intact (having no exons under P/A variation), E-PAV (having at least one, but not all, exons missing in at least one accession), and CDS-PAV (having the entire CDS missing in at least one accession) genes. 39

3.9 Genomic context for intact (having no exons under P/A variation), E-PAV (having at least one, but not all, exons missing in at least one accession), and CDS-PAV (having the entire CDS missing in at least one accession) genes. 40
3.10 Distribution of Z-scores for standardised transcript abundance data in the affected accession. 42

4.1 Total transcript number influences alternative splicing level (ASL) detection but this bias can be corrected using a sampling method. 56
4.2 Variance in alternative splicing over evolutionary time. 57
4.3 Relationship between alternative splicing and organism complexity, assayed as cell type number. 58
4.4 Biplot of the first two principal components built from 13 functional genomic variables available for 24 species. 62
4.5 Distribution of both alternative splicing events ($n = 17,738$) and ESTs ($n = 646,634$) for the genome of *A. thaliana*, according to the expression level of the associated gene. 65
4.6 Distribution of both alternative splicing events ($n = 76,699$) and ESTs ($n = 4,510,520$) for the genome of *H. sapiens*, according to the expression level of the associated gene. 66
4.7 Distribution of both alternative splicing events ($n = 45,628$) and ESTs ($n = 1,403,152$) for the genome of *M. musculus*, according to the expression level of the associated gene. 66
4.8 Distribution of both alternative splicing events ($n = 2128$) and ESTs ($n = 80,237$) for the genome of *C. elegans*, according to the expression level of the associated gene. 67

List of Tables

2.1 Correlates of dN/dS and NI in *A. thaliana*, after alignment against *A. lyrata*, *T. parvula*, or both. 12

2.2 Relationship between the average number of alternative splicing events per gene and the difference in evolutionary rate estimates before and after codon removal from the exon edges. 19

3.1 Characteristics of E-PAV and CDS-PAV genes compared to genes with all exons present in all accessions. 37

3.2 Age categories for orthologues of *A. thaliana* genes. 50

4.1 Association between CTN and genomic features before and after phylogenetic signal correction in 24 eukaryotic species. 59

4.2 Forward stepwise regression analysis using 13 functional genomic variables as predictors of CTN. 60

4.3 Regression coefficients of various functional parameters with number of ESTs as independent variable (quadratic polynomial regression). 63

4.4 Regression coefficients of 14 genomic parameters with CTN as the dependent variable, using as estimates for each variable the residuals of a linear regression between variable x and the log-transformed number of ESTs per species. 64

List of Supplementary Tables

Tables too large to feature in the main body of this thesis are included on the attached CD.

Table S2.1. Relationship between dN/dS and evolutionary rate predictors in *A. thaliana*

Table S2.2. Relationship between NI and evolutionary rate predictors in *A. thaliana*

Table S2.3. Partial correlations between dN/dS and evolutionary rate predictors in *A. thaliana*, controlling for expression level

Table S2.4. Average estimates of four evolutionary rate variables after sequential codon removal from the exon edges *vs.* random codon removal

Table S2.5. Characteristics of the dN/dS and NI distributions for dataset A (pairwise alignment of *A. thaliana* with *A. lyrata*) and dataset B (*A. thaliana* with *T. parvula*), before and after codon removal at the exon edges

Table S2.6. Estimates of four evolutionary rate variables for 3213 *A. thaliana* genes in dataset A (pairwise alignment of *A. thaliana* against *A. lyrata*) after having up to 10 codons removed from the exon edges

Table S2.7. Estimates of four evolutionary rate variables for 2041 *A. thaliana* genes in dataset A (pairwise alignment of *A. thaliana* against *A. lyrata*) after having up to 20 codons removed from the exon edges

Table S2.8. Estimates of four evolutionary rate variables for 1443 *A. thaliana* genes in dataset A (pairwise alignment of *A. thaliana* against *A. lyrata*) after having up to 30 codons removed from the exon edges

Table S2.9. Correlations between four selection strength/direction variables and 25 genomic characteristics, after sequential codon removal of 10 codons from exon edges *vs.* random codon removal

Table S2.10. Correlations between four selection strength/direction variables and 25 genomic characteristics, after sequential codon removal of 20 codons from exon edges *vs.* random codon removal

Table S2.11. Correlations between four selection strength/direction variables and 25 genomic characteristics, after sequential codon removal of 30 codons from exon edges *vs.* random codon removal

Table S2.12. Are estimates of *rho* for the correlation of each evolutionary rate variable with the genomic feature variable significantly different after sequential, compared to random, codon removal?

Table S2.13. Estimates of four evolutionary rate variables for 779 *A. thaliana* genes in dataset B (pairwise alignment of *A. thaliana* against *T. parvula*) after having up to 10 codons are removed from exon-intron junctions

Table S2.14. Estimates of four evolutionary rate variables for 350 *A. thaliana* genes in dataset B (pairwise alignment of *A. thaliana* against *T. parvula*) after having up to 20 codons are removed from exon-intron junctions

Table S2.15. Estimates of four evolutionary rate variables for 174 *A. thaliana* genes in dataset B (pairwise alignment of *A. thaliana* against *T. parvula*) after having up to 30 codons are removed from exon-intron junctions

Table S2.16. dN/dS estimates using codons common to the alignment of *A. thaliana*, *A. lyrata* and *T. parvula*

Table S2.17. Relationship between dN/dS and evolutionary rate predictors using estimates derived from codons common to the alignment of *A. thaliana*, *A. lyrata* and *T. parvula*

Table S2.18. Estimates of four evolutionary rate variables for 73 *A. thaliana* genes in dataset C (multiple sequence alignment of *A. thaliana* against *A. lyrata* and *T. parvula*) after having up to 10 codons are removed from exon-intron junctions

Table S3.1. Exons missing in 17 *A. thaliana* accessions, relative to Col-0

Table S3.2. Genome deletions due to missing exons in 17 *A. thaliana* accessions, relative to Col-0

Table S3.3. Transposable element density in the sequence adjacent to E-PAV genes

Table S3.4. Transposable element density in the sequence adjacent to CDS-PAV genes

Table S3.5. *A. thaliana* genes showing accession-specific exon loss

Table S3.6. Characteristics of GOslim categories and families enriched in genes with exon presence/absence variation

Table S4.1. Functional genomic parameter data, organism complexity and sources of genome annotations per species

Table S4.2. Regression coefficients of genomic parameters with CTN as the dependent variable

Table S4.3. Regression coefficients of genomic parameters with CTN as the dependent variable in the fungi-metazoan lineage only

Table S4.4. Correlation matrix (Spearman's *rho*) for 13 functional genomic variables in 24 species

Table S4.5. Correlation matrix (Spearman's *rho*) for 13 functional genomic variables in 15 fungi-metazoan species

Table S4.6. Regression coefficients of 14 genomic parameters with CTN as the dependent variable, using as estimates for each variable the residuals of a quadratic polynomial regression between variable x and the number of ESTs per species

Table S4.7. Regression coefficients of 14 genomic parameters with CTN as the dependent variable, using as estimates for each variable the residuals of a linear regression between variable *x* and the log-transformed number of ESTs per species

Table S4.8. Regression coefficients of genomic parameters with the average CTN of each species in the same order as the dependent variable

Table S4.9. Regression coefficients of genomic parameters with the average CTN of each species in the same order as the dependent variable, in the fungi-metazoan lineage only

Table S4.10. Relationship between the average number of alternative splicing events per gene and % of PPI domains per gene

Table S4.11. Estimates of cell type number

Abbreviations

ADA - *Arabidopsis* development atlas

ASE - alternative splicing event

ASL - alternative splicing level (average number of alternative splicing events per gene)

ASP - alternative splicing prevalence (proportion of alternatively spliced genes out of the total analysed)

BLAST - basic local alignment search tool

CC - coiled-coil

CDS - coding sequence

CNL - CC-NBS-LRR

CTN - cell type number

dN - number of non-synonymous substitutions per non-synonymous site

D_n - number of non-silent substitutions

DNA - deoxyribonucleic acid

dS - number of synonymous substitutions per synonymous site

D_s - number of silent substitutions

ENC - effective number of codons

ESE - exonic splice enhancer

EST - expressed sequence tag

F_{op} - frequency of optimal codons

gcRMA - robust multi-array analysis corrected for the GC-content of the oligo

GMAP - genomic mapping and alignment program

GO - gene ontology

LINE - long interspersed nuclear element

LRR - leucine-rich repeat

MPSS - massively parallel signature sequencing

mRNA - messenger RNA (ribonucleic acid)

NB-ARC - a nucleotide-binding adaptor shared by APAF-1 (apoptotic protease activating factor 1), R (resistance) genes and CED-4 (cell death protein 4)

NBS - nucleotide binding site

N_e - effective population size

NI - neutrality index

PAML - phylogenetic analysis by maximum likelihood

PAV - presence/absence variation

PC - principal component

PGLS - phylogenetic generalized least squares

P_n - number of non-silent polymorphisms

P_s - number of silent polymorphisms

PPI - protein-protein interaction

PRANK - probabilistic alignment kit

***R* gene** - resistance gene

RNA-seq - RNA (ribonucleic acid) sequencing

SINE - short interspersed nuclear element

SNP - single nucleotide polymorphism

TAIR - the *Arabidopsis* information resource

TE - transposable element

TIR - Toll/Interleukin-1 receptor

TNL - TIR-NBS-LRR

Chapter 1

Introduction

Since life first appeared on Earth, organisms have constantly changed, diversifying to populate most environments. The evolution of species proceeds as a result of the accumulation of changes in the prevalence of heritable genetic variations resulting from damage to, or errors during the replication of, DNA. What drives the observed changes in allele frequencies in different populations has been a matter of intense research for decades. Generally, changes in allele frequencies result from two main forces: selection and drift. Drift introduces noise into a population's distribution of allele frequencies by the random sampling of gametes in each generation, whereas selection refers to variations in allele frequencies as a result of their effect on fitness, either through increasing the carrier's reproductive success or its survival chances. At the molecular level, selection can act at individual genomic positions in a number of ways [1, 2]. Heritable variants (i.e. alleles) whose phenotypic expression increases fitness will increase in frequency over time, eventually replacing the ancestral variant from the population and becoming 'fixed'. This process is known either as positive or directional selection, or adaptation, as the fitness of an organism is adjusted with regard to its environment. By contrast, variation that decreases fitness will reduce in frequency within a population over time, a process known as negative or purifying selection. As new mutations have a higher probability of being deleterious than beneficial, more sites in the genome are under purifying than positive selection [3]. Finally, if the fitness advantage conferred by a particular variant is restricted to hetero-, rather than homozygotes, then the variant is maintained indefinitely within the population, a process known as overdominance. This is a type of balancing selection, a term encompassing various selective regimes that can stably maintain polymorphisms in a population [4]. It is also necessary to note that selection acts only on variation which impacts upon fitness, i.e. variation that affects the phenotype. Many mutations, however, do no such thing – they are selectively neutral and as such their fate is determined by drift. Kimura [5] proposed that variation at the molecular level is not, for the most part, related to fitness. Known as the 'neutral theory of evolution', this represents one of the most important concepts in molecular evolution.

Although empirical evidence in support of this theory suggests a quantitative dominance

of neutral processes in genome evolution [6] – that more mutations are fixed by drift than by selection because the majority of fixed mutations are selectively neutral – there nevertheless remains considerable debate as to the relative weight of positive selection and genetic drift in driving mutations to fixation, with 'neutralism' and 'selectionism' the extremes of an explanatory spectrum for describing patterns of whole-genome variation [7].

It is generally accepted that some types of variant are more likely to be under selection – for instance, point mutations are considered less likely to have deleterious effects than frame-shifting insertions or deletions, as are mutations within intergenic regions, which are less likely to have impact upon the phenotype than changes to gene sequences. Furthermore, within the coding regions of genes, synonymous sites – where base substitutions do not alter the amino acid encoded by the corresponding codon – are thought to be largely free of selective constraints compared to non-synonymous sites [8, 9].

This difference between synonymous and non-synonymous sites has been extensively used to quantify the extent of selective constraint acting on any given gene. The rate of non-synonymous substitutions per non-synonymous site (dN) can be used as an indicator of positive or purifying selection. The rate of synonymous substitutions per synonymous site (dS) can be taken as an indicator of the background rate of mutations and is commonly used to correct dN, producing a ratio, dN/dS. If drift alone is the force determining which mutations in a sequence reach fixation (i.e. the sequence is evolving neutrally), the expectation is that dN/dS approximates 1. The dN/dS ratio will exceed unity if natural selection promotes changes to the protein sequence [10], in accordance with theoretical work on various population genetic models, such as that of Wright-Fisher and Moran [11]. Deviations from unity can then be interpreted as the action of selection – a high dN is thought to stem primarily from gene-specific selective pressures related to the functionality of their protein products, and implies, although is not a definitive signature of, positive selection [12]. Statistically significant deviations from a neutral expectation can thus be used to infer the strength and direction of selection acting upon a given sequence, and in this respect, it is now commonplace to identify, and then annotate, putatively selected loci on a whole-genome basis [13, 14].

dN/dS ratios have been calculated on a genomic scale to identify genes which are undergoing positive selection in the hope that these provide clues as to the phenotypic adaptations distinguishing any two species or to reveal fundamental relationships between the characteristics and activity of genes, and their associated patterns of sequence evolution. Studies examining a wide variety of taxa have found a number of relationships. For instance, species-specific genes are less likely to be essential than genes present in ancestral species [15, 16], while more highly and broadly expressed are often more highly conserved [17]. In addition, a gene's age is related to the level of selective constraint it experiences: newer genes are less likely to be essential and accordingly are often under comparatively relaxed purifying selection [18, 19], i.e. would be more likely to have a

higher dN/dS ratio. However, as gene age is estimated by the existence of orthologues at different phylogenetic depths, this can confound the relationship – a sequence with a dN/dS ratio at or exceeding unity will by definition have relatively weaker conservation and lower sequence similarity to any orthologue. The relationship of age to evolutionary rate is arguably an artefact arising from the inability to detect homology across large distances [20]. There are a number of additional factors which need to be considered when establishing the existence and strength of these associations.

First, many earlier studies were based on a comparison of two species which makes it impossible to assess the directionality of any substitutions. In these cases, dN/dS ratios represent a composite of changes in the two lineages compared, and as such an increasing number of studies resolve the issue by using an outgroup species to estimate lineage-specific dN/dS [21, 22, 23, 24, 25].

Secondly, the use of dS as a denominator to correct dN for the background mutation rate assumes that mutations at synonymous sites are indeed under neutral evolution. This is not always the case. In addition to the biasing effects of preferential codon usage, whereby synonymous codons are used unevenly amongst the set of genes in a genome [26, 27] and GC-biased gene conversion, whereby GC in AT/GC heterozygotes is preferentially fixated and can be falsely interpreted as positive selection [28, 29], it has been shown that sequences flanking introns show higher conservation both at synonymous and non-synonymous sites as they harbour exonic splice enhancers [30], conserved sequence motifs that facilitate the assembly of splicing complexes [31, 32, 33]. How these factors influence the associations between sequence evolution and gene characteristics is unknown.

In addition, multiple – and contradictory – interpretations can be placed upon a dN/dS ratio (or that of any other selection test), as these estimates, in themselves, do not explicitly rule out alternative hypotheses [34]. Consider, for instance, if a sequence is found to have a 'high' dN/dS ratio (one exceeding unity) – as established above, this is consistent with a scenario of positive selection, but is also consistent with a scenario of comparatively relaxed purifying selection across the sequence, or of having fewer sites undergoing selection. A hard threshold that distinguishes between these two opposing selective forces is unreasonable and as such, to assist in interpretation, contextual information is necessary.

With the increasing availability of multiple genomes from a single species, it is now possible to take into account the rate of polymorphisms within a species at each genomic site. This allows the calculation of neutrality indices, measures of the degree and direction of departure of a sequence from a neutral expectation which take into account the numbers both of silent and non-silent polymorphisms and substitutions [35]. The classical neutrality index, the McDonald-Kreitman test, uses these estimates to compute the fraction of substitutions at functional sites that were driven to fixation by positive selection [36]. This test has the null hypothesis of neutrality, i.e. that the ratios of intra- and inter-species non-synonymous to synonymous variation are equal. Positive selection is

inferred when inter-species exceeds intra-species variation – adaptive mutations spread throughout a population rapidly and so affect the number of observed substitutions (i.e. divergence), but not the number of polymorphisms [37].

Genomic data is now available for multiple individuals of certain species, including humans [38] and the model plant *Arabidopsis thaliana* [39, 40], which has allowed exploration of other aspects of intraspecies variation. In particular, the discovery of partial or whole gene polymorphic deletions, affecting a significant proportion of the gene pool, has received increasing attention. A number of authors have proposed that polymorphic partial or whole gene deletions could play an important role in explaining the observed phenotypic variation between individuals of the same species. This is particularly notable for disease resistance phenotypes in *A. thaliana*, as *R* (resistance) genes show significant sequence divergence [41, 42] as well as enrichment for polymorphic deletions [43]. Against this selectionist model, however, polymorphic gene deletions can be explained using a neutral hypothesis whereby those genes under relaxed constraints would be the ones most likely to be polymorphically lost. No systematic test to distinguish between the two hypotheses has yet been carried out.

All observed phenotypic adaptations are, by definition, ultimately encoded in the genome. Untangling the relationships between specific phenotypes and changes at the genomic level has proven to be a difficult task, complicated primarily by the fact that at the molecular level, most changes are now known to be associated not with selection but neutral processes. These may affect the interpretation of any relationships. For instance, although particular changes at the molecular level may be attributed to selection, the efficacy of selection - i.e. the probability of fixation - is dependent on the effective population size (N_e) of the species given the relative fitness of the phenotype, expressed as the selection coefficient s [44]. s is considered the proportion by which the phenotype of interest is less fit, in terms of fertile progeny, and the rate of adaptive evolution proportional to $N_e s$ if $N_e s >> 1$ [45]. In general, the efficiency of purifying selection decreases concomitant with reductions in N_e (as proportionately fewer mutations have $N_e s >> 1$), resulting in a relative increase in the strength of drift and the increased accumulation of slightly deleterious mutations [46, 47, 48]. It is also of note that as more closely related species tend to increasingly share phenotypic characteristics, then any comparisons of genomic changes between species (irrespective of the evolutionary forces acting upon them) may in part be explained by their degree of phylogenetic relatedness.

It is thus of interest to compile and compare evidence of selection against evidence of neutral processes as acting across any given genome.

So, how do we distinguish the effects of selection from neutral processes? In this thesis, I address this question using a comparative genomics approach, taking advantage of the increasing availability of high-throughput sequencing data. In each chapter I address the adaptive significance of three different aspects of genome evolution: (i) rates of protein evolution (dN/dS), (ii) presence/absence variation (PAV; intraspecies structural variation

in a gene, in full or in part, such that genic sequence is present in some individuals of a species but not others), and (iii) the evolution of alternative splicing (a post-transcriptional process by which multiple distinct transcripts are produced from a single gene) across eukaryotes.

These represent test cases in which possible adaptive interpretations of explicitly genome-wide phenomena can be contrasted with explanations that primarily invoke neutral processes. In two of the chapters, I focus on the model plant *A. thaliana*, of particular note in that it is a near obligate selfer, having a patchy distribution of inbred populations with relatively rare outcrossed matings between different ecotypes [49]. Selfing increases genome-wide homozygosity, and thus decreases the number of gametes which may be independently sampled in a given population, in effect reducing N_e [50]. As a consequence, the efficacy of selection - particularly purifying selection - at purging weakly deleterious mutations is reduced [51]. By contrast, selfing species may more readily purge recessive deleterious mutations, 'exposing' them to selection by pairing them in a homozygote [52]. Arguably, however, the reduction in selective efficacy brought about by a reduced N_e are stronger than this 'exposing' effect [53], with studies reporting a general trend of relaxed selection in *A. thaliana* compared to outcrossing lineages, albeit with weak statistical support [54]. Nevertheless, notably reduced selective efficacy, in particular, has been identified in this species at silent sites [55].

In chapter 2, I address questions relating to the interpretation of dN/dS, a signature of selection, on a genome-wide basis. It is important to note that interpreting any estimates of the dN/dS ratio rely upon assumptions which may be violated. In particular, it is assumed the selective pressures acting over synonymous sites are constant such that they can be taken as a proxy of the mutation rate. This is not necessarily the case. For instance, differential patterns of codon usage have been characterised at mammalian exon-intron junctions [56, 57], consistent with enhanced conservation of synonymous sites in these regions so as to maintain accurate splicing. These in turn can affect the overall dN/dS estimate per gene [58, 59] and as such may mask the degree and direction of any selection occurring. This potential bias in dN/dS interpretation is addressed in the model plant *A. thaliana*. This is because the current consensus indicates that, in general, plant genomes are predominantly under purifying selection [44], with low estimates of the number of positively selected genes driving the divergence of sorghum [60] and maize [61], as well as both *A. thaliana* [62, 63] and *A. lyrata* [64]. It is possible that the failure to identify positively selected genes reflects a mostly neutral evolutionary process, or alternatively that other functional characteristics of genes have masked selection upon non-synonymous mutations. Given the potential effect of exon-intron junction conservation and that in *A. thaliana*, 75% of the genes are multi-exonic and 29% of the exons are short (< 100bp), per-gene estimates of dN/dS may thus be affected by a disproportionately high dS.

This chapter shows that higher conservation at exon-intron junctions in *A. thaliana* is an important component of dN/dS estimates by disproportionately increasing the propor-

tion of conserved synonymous sites. This partially accounts for the relationship between dN/dS and the functional characteristics of a gene and thus provides additional contextual information by which dN/dS estimates can be interpreted.

In chapter 3, I consider questions relating to presence/absence variation in the model plant *A. thaliana*. This chapter takes advantage of the recent genome sequencing of multiple *A. thaliana* accessions [39, 40], which identify extensive intraspecies variation in both gene and exon content [43], consistent with that observed in other plant species: maize [65], sorghum [66] and soybean [67]. As genes showing such structural variations are enriched for those involved in disease resistance, a function known to be fast evolving [68], this aspect of *A. thaliana* genome evolution suggests a strong contribution made by selection [43]. As previously explained, an adaptive explanation for a particular phenomenon necessitates that any observed variation has an effect upon the phenotype, such that it may influence fitness. In the case of genes involved in disease resistance, it is reasonable to suspect that the observed variation (PAV) leads to a phenotype visible to selection (novel protein structure), which may benefit fitness (enhanced disease resistance in response to a local pathogenic challenge).

Nevertheless, this adaptive hypothesis, although a compelling interpretation, has not been explicitly tested. This chapter contrasts selective against neutral explanations of this phenomenon, employing both selection tests and contextual information relating to gene structure, function, genomic location and essentiality, concluding that although adaptive scenarios cannot be explicitly discarded, PAV events can be explained without invoking them.

In chapter 4, I consider questions relating to the evolution of alternative splicing, a mechanism by which a single pre-mRNA can yield multiple functionally distinct mRNAs, potentially increasing proteome diversity at a level disproportionate to increases in gene number [69, 70]. Although a common process in 'higher' eukaryotes – particularly metazoans – increases in the proportion of genes undergoing alternative splicing [71] are known to correlate with decreases in the effective population size of species [72, 73]. This predicts an increase in the relative strength of drift to selection and the subsequent accumulation of slightly deleterious mutations, including those affecting splicing regulation. This would result in a higher number of transcripts, but not necessarily those that are functionally meaningful. As a consequence, many observed alternative splicing events are arguably selectively neutral and that their existence is as a transient state, prior to being fixed, or lost, due to drift. As such, the observed increases in alternative splicing for eukaryotes may primarily be non-adaptive in nature, and that – in general – a low coding potential for alternatively spliced transcripts is expected. It has thus been argued that the effect of genetic drift, a neutral process, has a dominant role in the evolution of alternative splicing and that the majority of observed alternative splicing events are, rather than adaptive, 'transcriptome noise' [74, 75, 76, 77].

Nevertheless, there is an accumulating body of evidence to suggest that alternative splic-

ing has had a central role in many biological processes. For instance, it is known that alternative splicing can lead to physiologically meaningful changes in the domain content, binding properties and stability of a protein, along with its intracellular localization and enzymatic activity [78, 79, 80], that genes whose protein products are involved in protein-protein interactions have higher levels of alternative splicing [81], that the alternatively spliced regions of genes often comprise interaction sites for proteins and their binding partners [82], and that splicing plays a role in numerous biological processes, such as the virulence of pathogenic fungi [83], neuronal regulation in both primates [84, 85] and rats [86], adaptive immune responses in invertebrates [87, 88], sex determination in *Drosophila* [89, 90], and the differential morphological development of caste-specific individuals in eusocial ants [91] and honey bees [92]. Genes with high levels of alternative splicing are also enriched for cytoskeleton-associated functions in vertebrates [93], suggesting a potential role in the evolution of cellular complexity.

As such, although the above studies, in sum, suggest numerous possibilities for functionally relevant splicing, these observations can nevertheless be reconciled by suggesting a general trend of non-functional splicing is the norm.

In the absence of experimental verification, the functionality of splicing events is hard to assess and the problem is often approached using proxies. For instance, indicative of potential functionality is that an alternative exon is conserved across species. That proportionately few have been observed between, e.g., humans and mice, suggests that splicing events are not, in general, of phenotypic significance ([94], but see also [95]), and that the distribution of alternative splice sites is instead consistent with a neutral model of random fixation [96]. Despite this, conserved patterns of differential exon usage have also been observed in primates, albeit for a 'sizeable minority' of genes [97]. It is still unclear, however, as to what extent 'the majority' of a genome's splicing is noisy and how sizeable is 'the minority' of splicing that is not.

Overall, it is reasonable to expect the evolution of alternative splicing to have involved the interplay of both adaptive and neutral processes, although it is unknown as to which evolutionary force is of most significance in shaping the observed distributions of alternative splicing events.

The main obstacle to comparative studies of alternative splicing is that differences in transcript coverage between species can distort the proportion of genes classified as undergoing alternative splicing and the number of splicing events detected [98, 71, 99]. This chapter details the computation of a comparable index of alternative splicing, spanning all major eukaryotic taxa and correcting for this bias via the method of [98], and establishes that alternative splicing is a strong predictor of organism complexity. Importantly, we found no evidence to suggest that this relationship is explained by drift due to a reduced effective population size in 'more complex' species. These results are in principle consistent with an adaptive role of alternative splicing in determining a genome's functional information capacity, by facilitating increased transcript diversification in 'more

complex' species. As such, this chapter contributes to the ongoing debate surrounding the contributions made by selection and drift towards alternative splicing evolution.

In sum, this thesis addresses the means by which the activity of selection at the genome-wide level can be distinguished from neutral processes.

Although the findings address the relative activity of selection and neutral forces upon specific biological processes, they may also be placed in a broader context to illustrate a more general conclusion. This argues that given the rise in high-throughput sequencing data, it is no longer wholly sufficient to determine the action of selection by taking the results of 'selection tests' (such as dN/dS or a neutrality index) in isolation. Rather, as it is viable to consider large-scale adaptive phenomena acting across genomes, it is arguably necessary to consider multiple signatures of selection simultaneously and, by placing these results in the wider context of what is known of genome-wide evolutionary processes, to construct a biologically meaningful narrative for these observations.

Chapter 2

Lineage-specific sequence evolution and exon edge conservation partially explain the relationship of evolutionary rate and expression level in *A. thaliana*

2.1 Summary

Rapidly evolving proteins can aid the identification of genes underlying phenotypic adaptation across taxa, but functional and structural elements of genes can also affect evolutionary rates. In plants, the 'edges' of exons, flanking intron junctions, are known to contain splice enhancers and to have a higher degree of conservation compared to the remainder of the coding region [100]. However, the extent to which these regions, if at all, may be masking indicators of positive selection or account for the relationship between dN/dS and other genomic parameters is unclear. We investigate the effects of exon edge conservation on the relationship of dN/dS to various sequence characteristics and gene expression parameters using the model plant *Arabidopsis thaliana*. We also obtain lineage-specific dN/dS estimates, making use of the recently sequenced genome *Thellungiella parvula*, the second closest sequenced relative after the sister species *Arabidopsis lyrata*. Overall, we find that the effect of exon edge conservation, as well as the use of lineage-specific substitution estimates, upon overall dN/dS partly explains the relationship between the rates of protein evolution and expression level, as well as supporting a stronger link between dN/dS and gene length. Furthermore, the removal of exon edges shifts dN/dS estimates upwards, increasing the proportion of genes potentially under adaptive selection. We conclude that lineage-specific substitutions and exon edge conservation have an important effect on dN/dS ratios and should be considered when assessing their relationship with other genomic parameters.

2.2 Introduction

The rate of sequence evolution is known to vary between genes, particularly at non-synonymous sites [101], with variations in the non-synonymous rate of substitution (dN) thought to stem primarily from gene-specific selective pressures related to the functionality of their protein products [12]. As such, a ratio of non-synonymous to synonymous substitutions (dN/dS) is often used to identify those genes likely to be involved in adaptation [102]. Determining which genes are under selection, in what manner, and how this varies between species, is important for understanding how genetic diversity is maintained and the relative importance of opposing selective forces in shaping a species' genome.

The increased availability of sequenced genomes – from six sequenced species in 1997 [103] to thousands today [104] – has allowed genome-wide patterns of selection to be investigated. Among them, the model plant *Arabidopsis thaliana*, one of the better characterised plant species, is an ideal organism to investigate genome-wide signatures of selection as multiple genomes of this species have been sequenced [40, 39], along with the genomes of its sister species *Arabidopsis lyrata* [105] (from which *A. thaliana* diverged approx. 13 million years ago [106]) and a more distant relative, the extremophile crucifer *Thellungiella parvula* (from which *A. thaliana* diverged approx. 43 million years ago) [107]. As the available data regards intraspecies diversity as well as species divergence, this is a substantial resource for analyses of genome-wide evolutionary rate variation [63, 108].

The current consensus indicates that plant genomes are predominantly under purifying selection [44], with low estimates for the number of positively selected genes driving the divergence of sorghum [60] and maize [61], as well as both *A. thaliana* [62, 63] and *A. lyrata* [64]. It is possible that the failure to identify positively selected genes reflects a mostly neutral evolutionary process, particularly in plants, or alternatively that other functional characteristics of genes have masked selection upon non-synonymous mutations.

When comparing only two species, dN/dS ratios are a composite of substitutions occurring in either lineage. Since genomic parameters are only assessed in one species, i.e. in *A. thaliana* but not *A. lyrata*, then any relationship between these and dN/dS may be diluted. As such, it is important to calculate lineage-specific dN/dS by using an outgroup species (e.g. [21, 22, 23, 24, 25]). In this respect, the genome of *Thellungiella parvula* [107] provides an outgroup suitable for assessing lineage-specific sequence evolution, potentially mitigating bias in dN/dS.

Interpretation of dN/dS ratios assumes that synonymous substitutions are mostly evolving under neutral conditions or that, at the very least, the selective pressures acting over these sites are constant such that they can be taken as a proxy of the mutation rate. However, exon sequences can contain exonic splicing enhancers (ESEs), sequence motifs involved in both constitutive and regulated splicing by facilitating the assembly of splicing complexes [33, 31, 32]. In humans, 6-8bp ESEs are enriched in the vicinity of splice sites,

in particular 80-120 bases downstream of a splice acceptor [109], with these sequences under increased selection in intron-rich genomes [110]. As higher conservation at the edges of exons, particularly at synonymous sites, can reflect differential patterns of codon usage [56, 57] and affect the overall dN/dS estimate per gene [58, 59], this can influence any reported relationship between dN/dS and various genomic parameters, particularly in compact, intron-rich genomes. In *A. thaliana*'s genome, 75% of the genes are multi-exonic, 29% of the exons are below 100bp, the median exon length is 53 codons, and ESE hexamers have been identified [100]. Thus, ESE conservation may have a strong impact on estimates of dN/dS, and as such, estimates of the relative contribution of positive and purifying selection to *A. thaliana* genome evolution.

In a substantial number of species (e.g. *A. thaliana*, *B. subtilis*, *C. elegans*, *D. melanogaster*, *E. coli*, *H. sapiens*, *M. musculus*, the legume *M. trunculata*, *S. cerevisiae*, *S. typhimurium*, *T. rubripes* and the viral order Mononegavirales) expression level has been found to be the best predictor of dN/dS ratios [111, 112, 113, 114, 115, 116, 117, 17, 118, 119, 120, 121]. Additional variables found to be associated with dN/dS include expression breadth (an estimate of the proportion of tissues in which a gene is expressed) [122, 123, 124, 125], codon usage bias [126, 127, 128, 129, 130, 131], GC content [132, 133], protein multi-functionality [134, 135], the number of interacting partners per protein [136, 137, 138, 139, 140], recombination rate [141, 142], gene/protein length [143, 35, 144, 145] and both intron number and length [146, 147, 148].

It is not yet known how dN/dS estimates are influenced by exon edge conservation [59], nor whether this affects the indicators of positive selection within a genome or the covariance between the rate of sequence evolution and any given genomic parameter.

Here we address this issue by examining coding sequence evolution in *A. thaliana*, with *A. lyrata* and *T. parvula* as comparison species, to investigate: *(i)* whether selection to conserve exon-intron junctions has a significant effect on dN/dS estimates and if so whether this affects *(ii)* the relationship between expression level and evolutionary rate, *(iii)* the relationship between other structural and functional parameters previously identified as dN/dS correlates in one or more other species, and *(iv)* whether accounting for exon edge conservation and lineage-specific dN/dS may unmask higher levels of positive selection.

2.3 Results

2.3.1 Correlates of dN/dS in *A. thaliana*

We found a significant correlation between expression level estimates based on RNA-seq data and dN/dS estimated using *A. thaliana-A. lyrata* alignments ($rho = -0.42$, $p < 2.2 \times 10^{-16}$, $n = 21,198$; Tables 2.1 and S2.1). Similar results were obtained when using independent expression data from two alternative platforms, microarrays and MPSS, as well as when applying four normalisation procedures previously used for each set of

estimates (Tables 2.1 and S2.1). All estimates of gene expression level were also significantly correlated with NI (Tables 2.1 and S2.2). As similar results were obtained when using all gene expression data sources, all analyses involving expression level shown below (unless otherwise stated) refer to the RNA-seq data [39].

Variable	Alignments of *A. thaliana* with *A. lyrata*		Alignments of *A. thaliana* with *T. parvula*		Alignments of *A. thaliana* with both *A. lyrata* and *T. parvula*	
	dN/dS	NI	dN/dS	NI	dN/dS	NI
Average exon length	0.103	-0.026	0.017 (*)	0.045	-0.161	-0.043
Average intron length	-0.070	0.043	-0.052	0.061	-0.012 (*)	0.088
Gene length	-0.243	0.092	-0.067	-0.047	-0.168	0.044
Primary transcript length	-0.243	0.092	-0.067	-0.047	-0.168	0.043
Protein length	-0.124	0.050	-0.015 (*)	-0.060	-0.197	-0.034
Total exon length	-0.203	0.075	-0.066	-0.039	-0.212	0.005 (*)
Total intron length	-0.228	0.086	-0.056	-0.041	0.001 (*)	0.089
UTR length (5')	-0.183	0.032	-0.131	0.003 (*)	-0.060	0.035
UTR length (3')	-0.122	0.053	-0.070	0.040	-0.037	0.086
Expression breadth	-0.399	0.120	-0.284	0.117	-0.148	0.232
Exp. level (RNA-seq)	-0.415	0.145	-0.285	0.117	-0.179	0.217
Protein abundance	-0.302	0.078	-0.241	0.095	-0.116	0.194
Tau	0.277	-0.088	0.210	-0.092	0.133	-0.175
Effective number of codons	0.059	-0.016	0.065	-0.035	0.057	-0.043
Frequency of optimal codons	-0.194	0.065	-0.187	0.116	-0.114	0.176
GC (%)	-0.009 (*)	0.036	-0.057	0.081	-0.160	0.038
Intron density	-0.158	0.048	-0.022	-0.052	0.048	0.064
Total no. of introns	-0.212	0.071	-0.038	-0.069	0.002 (*)	0.062
Multifunctionality	-0.132	-0.013 (*)	-0.137	$-1.18\text{x}10^{-4}$ (*)	-0.133	-0.012 (*)
Protein-protein interactions	-0.060	0.031	-0.084	0.069	-0.022 (*)	0.152
Recombination rate	0.007 (*)	-0.058	-0.011 (*)	-0.019 (*)	0.003 (*)	-0.022 (*)

Table 2.1: Correlates of dN/dS and NI in *A. thaliana*, after alignment against *A. lyrata*, *T. parvula*, or both.
Correlation strengths are shown as Spearman's *rho*. Statistically insignificant values ($p < 0.05$) are marked with an asterisk. An expanded version of this data, showing p-values and sample sizes, is available as Tables S2.1 and S2.2.

We then assessed the relationship between dN/dS and gene expression breadth (measured both as the proportion of tissues in which a gene is expressed and the tissue specificity index *tau* [149]), gene length and intron content, GC content, codon usage bias (measured as the frequency of optimal codons, F_{op}), protein multifunctionality and interactivity, as well as recombination rate. We observed significant associations between dN/dS and all parameters tested with the exception of GC content and recombination rate (Tables 2.1 and S2.1). Expression level was found to be the strongest predictor of dN/dS followed by expression breadth, gene length and intron content (Tables 2.1 and S2.1). In the case of NI, we found that expression level and expression breadth explain approximately equal proportions of the variance and that in each case these variables are the dominant predictors (Tables 2.1 and S2.2).

As many of the variables found to be significantly associated with dN/dS ratios are themselves covariates of expression level, it is possible that some parameters may co-vary with

dN/dS as a by-product of their relationship with expression level. Thus, to better understand the relative contribution of genomic features to dN/dS we assessed the relationship between individual parameters and dN/dS after controlling for the effect of expression level. To do this, we calculated partial Spearman's correlation coefficients, finding that all significant predictors of dN/dS remained so after correcting for their covariance with expression level (Table S2.3).

Overall, we observe that expression level is the strongest predictor of dN/dS. Furthermore, the partial correlation coefficients for each variable show that the correlation of most variables with dN/dS is only partially accounted for by their covariance with expression level.

2.3.2 Accounting for exon edge conservation influences dN/dS and its relationship with various genomic parameters, and unmasks higher levels of positive selection

Using pairwise alignments of *A. thaliana* with *A. lyrata*, we find that codon removal at the edges of exons results in increased dN, dS and dN/dS estimates when compared to those estimates calculated after random codon removal from any position in the coding sequence (Figure 2.1 and Table S2.4). Estimates of NI were found to decrease with codon removal compared to those NI values calculated from sequences where codons were removed at random positions, suggesting a weakening in the departure of sequence evolution from a neutral expectation (Figure 2.1 and Table S2.4). These patterns are consistent with exon edges being under selective constraint with these regions having comparatively fewer non-synonymous substitutions than sequence elsewhere in the gene. The above findings suggest that codons at the ends of exons have a strong effect on dN/dS values, reflecting the high levels of purifying selection at these positions acting on both synonymous and non-synonymous sites.

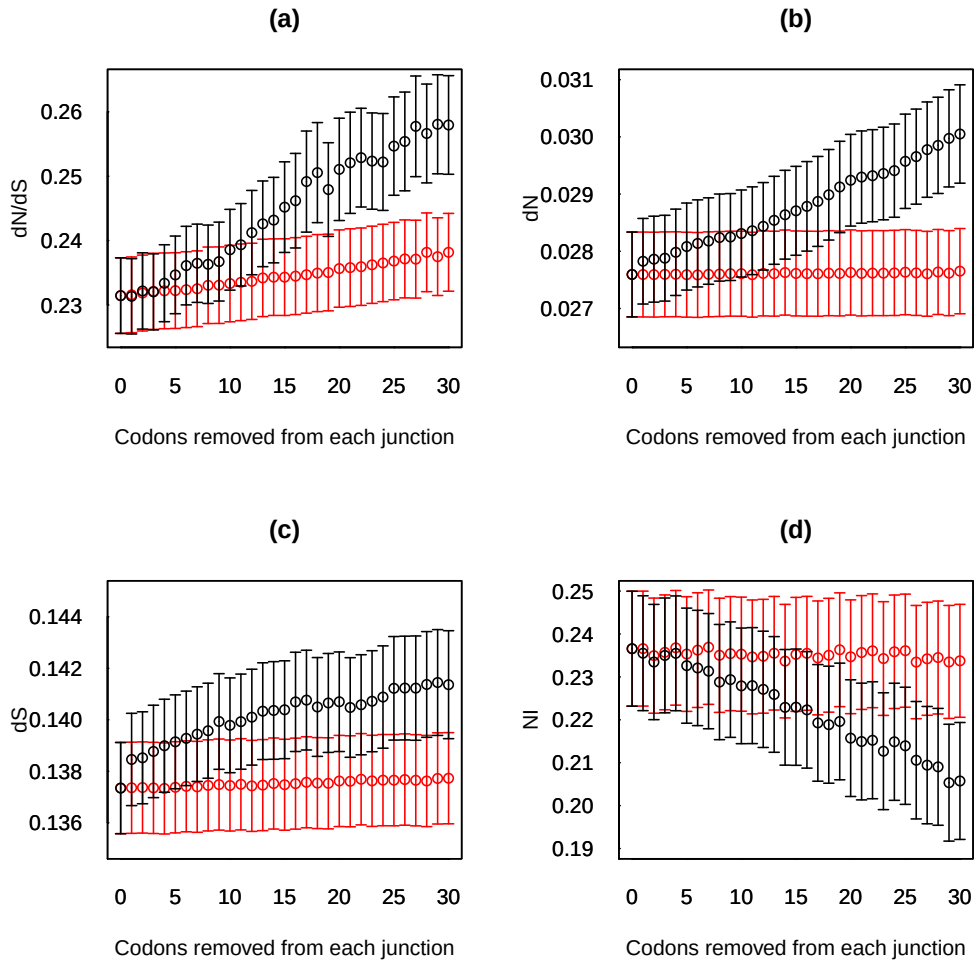

Figure 2.1: dN, dS, dN/dS and NI after exon edge removal.
dN/dS (a), dN, (b), dS (c) and NI (d) for a sample of 1443 genes with at least one fully alignable exon between *A. thaliana* and *A. lyrata*, after removing one codon at a time from exon edges (black), to a maximum of 30. The effects of random codon removal are shown in red. Statistical analyses are shown in Table S2.7. The above patterns are consistent with data obtained using a lower number of removed codons, up to a maximum of 10 (Table S2.5) or 20 (Table S2.6).

This results in a shift in the distribution of these estimates, suggesting that, in general, sequential codon removal leads to a higher dN/dS than random removal (paired Mann-Whitney U test, null hypothesis: location shift > 0, $p = 3.29 \times 10^{-5}$, $n = 1443$; Table S2.5). Complimentary results were obtained when using NI (paired Mann-Whitney U test, null hypothesis: location shift < 0, $p = 1.24 \times 10^{-16}$), and after the removal of 10 codons (dN/dS, $p = 9.58 \times 10^{-7}$; NI, $p = 9.90 \times 10^{-32}$; $n = 2041$) and 20 codons (dN/dS, $p = 5.73 \times 10^{-6}$; NI, $p = 2.68 \times 10^{-21}$; $n = 1443$). 35 genes (2.4% of the sample analysed) were found to have dN/dS values higher than 1, a potential indicator of positive selection, after the sequential removal of 30 codons from exon edges, compared to 20 genes after codon removal at random positions (χ^2 = 11.25, p = 7.96x10^{-4}; Table S2.5). Similar results were obtained after the removal of 20 codons (after sequential codon removal, 50 genes have $dN/dS > 1$; after random removal, 35 genes, χ^2= 6.43,

$p = 0.011$), and after the removal of 10 codons (after sequential codon removal, 77 genes have $dN/dS > 1$; after random removal, 58 genes; $\chi^2 = 6.22$, $p = 0.013$). In general, exon-intron junction removal shifts dN/dS values towards a range indicative of either stronger positive or relaxed purifying selection, with an overall increase in the proportion of genes potentially under adaptive selection.

In order to understand the effect of higher conservation at the exon edges on the relationships between dN/dS and other genomic parameters, we reanalysed the correlations after removing 10 ($n = 3213$ genes), 20 ($n = 2041$ genes) or 30 ($n = 1443$ genes) codons from the edge of each exon (Tables S2.4, S2.6, S2.7 and S2.8). We found that the correlation strength of dN/dS with numerous genomic features – in particular, expression level and expression breadth – decreased after the removal of exon edges. We observed only marginal changes to these correlation coefficients after removing an equivalent number of codons from random positions (Tables S2.9, S2.10, S2.11 and S2.12). This suggests that after the removal of exon edges, the decreased correlation strength between dN/dS and these genomic parameters is not explained by increased noisiness resulting from the use of shorter sequences to estimate dN/dS. The inference is also that a dN/dS based test of selection is most acute for more highly expressed genes and that stronger correlations of dN/dS with their various characteristics reflect the stronger constraints upon them. Furthermore, when considering NI, several variables including expression level, expression breadth, intron density and various measures of gene length become stronger predictors of NI (Table S2.12). These findings show that increased conservation at the edges of exons influences the relationship of dN/dS and NI to the structural and functional characteristics of a gene, suggesting a disproportionate pattern of substitutions in these regions that serves as a source of bias. Nevertheless, the relative order of these parameters as predictors of dN/dS remains largely unchanged with expression level still the dominant predictor.

2.3.3 Using the more distant relative *T. parvula* results in similar patterns to those found with comparisons to *A. lyrata*

When all analyses were repeated using dN/dS ratios estimated from alignments of *A. thaliana* and *T. parvula*, similar patterns to those found with *A.thaliana-A. lyrata* alignments were obtained. Expression level remains the strongest predictor of dN/dS , albeit with comparatively lower covariance ($rho = -0.29$, $p < 2.2 \times 10^{-16}$; Tables 2.1 and S2.1), with both expression breadth and codon usage bias explaining progressively smaller proportions of dN/dS variance. The relationship between NI and numerous genomic features showed similar patterns to those observed when using *A. thaliana-A. lyrata* alignments (Tables 2.1 and S2.2). Furthermore, the removal of exon edges from the *A. thaliana-T. parvula* alignments resulted in similarly reduced correlation strengths for dN/dS and numerous genomic features, but did not affect the relative dominance of expression as a predictor (Tables 2.1, S2.1, S2.4, S2.13, S2.14 and S2.15). Exon edge re-

moval also results in the significant shift of both dN/dS and NI distributions towards values indicative of relaxed selective constraint and/or stronger positive selection, although we find no increase in the proportion of genes with $dN/dS > 1$ (Table S2.5).

2.3.4 Reduced prominence of gene expression as a predictor of *A. thaliana*'s lineage-specific dN/dS

Lineage-specific dN/dS estimates were derived from multiple alignments of *A. thaliana* genes with *A. lyrata* and *T. parvula*, and revealed a marked decrease in the covariance between dN/dS, and various genomic parameters including expression level ($rho = 0.18$, $p < 2.2 \times 10^{-16}$, $n = 7086$) and expression breadth ($rho = -0.15$, $p < 2.2 \times 10^{-16}$, $n = 5758$), with total exon length becoming the strongest covariant of dN/dS ($rho = -0.21$, $p < 2.2 \times 10^{-16}$, $n = 7086$); Tables 2.1 and S2.1).

To rule out the possibility that reductions in both the absolute and relative strength of the correlation between dN/dS and gene expression when examining lineage-specific changes may be explained by differences in the gene/codon set tested, we re-calculated pairwise dN/dS for *A. thaliana* against *A. lyrata* and *T. parvula* using only those codons common to the multiple alignments of *A. thaliana*, *A. lyrata* and *T. parvula* used to estimate lineage-specific dN/dS (Table S2.16). We confirmed the marked reduction in the correlation strength between lineage-specific dN/dS and numerous genomic features when compared to pairwise comparisons of *A. thaliana* and either of *A. lyrata* or *T. parvula* (Table S2.17). Significant reductions in the correlation between dN/dS and expression level observed when examining lineage-specific dN/dS as contrasted to pairwise comparisons of *A. thaliana* with either *A. lyrata* or *T. parvula* were also confirmed when using the same codons in all cases. Statistical significance was tested using a T-test on the Z-transformed values of *rho*, as implemented by the paired.r method of the R package 'psych' [150] – $rho = -0.14$, -0.25 and -0.28, for lineage-specific, pairwise vs. *A. lyrata* and pairwise vs. *T. parvula* estimates of dN/dS respectively, with T-test $p = 2.38 \times 10^{-5}$ and 7.16×10^{-8} corresponding to significant reductions of the lineage-specific estimate from both the pairwise vs. *A. lyrata* estimate and the pairwise vs. *T. parvula*; Table S2.17). In contrast, the strength of the correlation between dN/dS and protein length is increased when lineage-specific dN/dS values were used instead of pairwise comparisons ($rho = -0.18$, -0.10 and -0.09, respectively, with T-test $p = 9.05 \times 10^{-3}$ and 9.42×10^{-4}; Table S2.17). Notably in this case, the strength of the correlation between protein length and *A. thaliana*'s lineage-specific dN/dS ($rho = -0.18$, $p = 4.70 \times 10^{-10}$; Table S2.17) was not weakened when controlling for the effect of expression level (partial $rho = -0.20$, $p = 3.47 \times 10^{-67}$, Table S2.3).

We further find that using lineage-specific substitution patterns markedly reduces the number of genes with $dN/dS > 1$ (21 genes have $dN/dS > 1$, 0.3% of the sample analysed) when compared to pairwise alignments of *A. thaliana* with *A. lyrata* (423 genes

have $dN/dS > 1$, 2% of the sample analysed; chi-sq. $p < 2.2 \times 10^{-16}$), but not when compared to alignments with *T. parvula* (41 genes have $dN/dS > 1$, 0.4% of the sample analysed, chi-sq. $p = 0.327$).

Taken together, these findings suggest that when examining lineage-specific dN/dS estimates, the prominence of gene expression is diminished with protein length becoming the dominant predictor. This pattern is not explained by variations in the sample of genes/codons used for the analyses. Importantly, we find no evidence that the use of lineage-specific dN/dS estimates unmasks any additional signatures of positive selection compared to pairwise alignments.

2.4 Discussion

2.4.1 Lineage-specific substitution estimates and the conservation of exon edges partially explain the association between gene expression and dN/dS in *A. thaliana*

Previous studies have shown that in mammalian species exonic splicing enhancer sequences, located nearer the ends of exons [151], result in higher conservation of synonymous sites at the exon edges, suggestive of selective constraint to maintain correct splicing [58, 59]. We show that, consistent with these findings, the removal of codons at the exon edges has a strong effect on the rate of substitutions at synonymous sites in *A. thaliana*, suggesting similar constraint, and associated functional importance, for ESE-containing regions. A more moderate increase was also observed in non-synonymous rates of substitutions reflecting that purifying selection at these sites is higher than the average observed at non-synonymous sites throughout the gene. When analysing estimates of dN/dS before and after the removal of codons at the exon edges we found a sharp increase, suggesting a disproportionate pattern of substitutions in these regions. Similar patterns were observed when analysing synonymous and non-synonymous polymorphisms with higher polymorphism rates per site found after the removal of exon edges. Consequently, the removal of exon edges resulted in a weaker association between dN/dS and measures of expression level and breadth. The relationship between dN/dS and other genomic parameters, particularly gene length, was also affected but to a lesser extent (Table S2.12). The relative prominence of different features was not affected after the removal of exon edges with gene expression remaining the strongest predictor. The observed decrease in the relationship of dN/dS and NI to gene expression after the removal of exon edges suggests that a stronger degree of purifying selection acting upon splice enhancer regions partly explains the association of dN/dS and NI to expression. From this we can infer stronger splice-mediated selection in more highly expressed genes.

Taking the above findings together, how does exon edge conservation relate to an increased emphasis upon gene length – particularly when lineage-specific dN/dS is consid-

ered – and a decreased emphasis on expression level as a predictor of evolutionary rate? These findings show that stronger constraint upon putative ESE-containing regions biases the per-gene estimate of dN/dS, but that this is partially masked by the stronger relationship dN/dS has with expression level. It is reasonable to ask if anything can explain this selective constraint in such a way as to also relate both to gene length and expression. One possible explanation may be the extent to which a gene is alternatively spliced. Alternative splicing has been shown to positively correlate with a gene's expression level [152] and its intron content [153]. As longer genes are more likely to have complex exon-intron architectures [154] this would increase, by probable combinations alone, the number of possible splicing events. If we assume that the exon edges are under increased selection for accurate alternative splicing compared to non-alternatively spliced exons, then those genes that have higher levels of alternative splicing are expected to show a greater discrepancy in evolutionary rate estimates before and after codon removal. Using estimates of the number of alternative splicing events per gene (see Materials and Methods), we find that dN/dS ratios (calculated from pairwise alignments of *A. thaliana* and *A. lyrata* to maximise sample size) are more strongly affected by codon removal from the exon edges in genes with higher levels of alternative splicing – for instance, the increase in dN after 10 codons are removed is significantly higher for genes with more splicing events ($rho = 0.13$, $p = 2.7 \times 10^{-4}$; Table 2.2). This pattern is maintained even when up to 30 codons are removed (Table 2.2). Although based upon a limited sample size, this finding merits further scrutiny as it shows that genes with alternative splicing events, compared to non-spliced genes, have a higher degree of conservation at exon edges relative to conservation of the remaining coding sequence.

Dataset	Evo. rate variable	No. of codons removed	*rho*	p	n
Pairwise alignment of *A. thaliana* with *A. lyrata*	dN	10	0.130	2.74×10^{-4}	781
		20	0.103	0.039	403
		30	0.133	0.040	239
	dS	10	0.041	0.385	462
		20	0.056	0.429	204
		30	0.095	0.292	124
	dN/dS	10	0.167	5.65×10^{-3}	273
		20	0.219	1.15×10^{-3}	217
		30	0.090	0.315	127
	NI	10	0.141	0.023	262
		20	0.072	0.278	226
		30	0.062	0.453	150
Pairwise alignment of *A. thaliana* with *T. parvula*	dN	10	0.049	0.468	219
		20	-0.052	0.631	88
		30	0.014	0.925	45
	dS	10	0.096	0.307	116
		20	0.035	0.826	42
		30	0.251	0.286	20
	dN/dS	10	0.011	0.916	92
		20	-0.083	0.569	49
		30	-0.185	0.311	32
	NI	10	0.052	0.609	98
		20	0.074	0.547	68
		30	0.083	0.613	40

Table 2.2: Relationship between the average number of alternative splicing events per gene and the difference in evolutionary rate estimates before and after codon removal from the exon edges.
The difference between dN, dS, dN/dS and NI estimates before and after codon removal) is a proxy for the degree of selective constraint on the junction region. Only those genes with statistically significant discrepancies in four evolutionary variables after sequential codon removal from exon-intron junctions, compared to random codon removal, are used in this analysis (summarised in Table S2.4; see Tables S2.5-S2.7, and S2.13-S2.15, for data). Statistically significant findings ($p < 0.05$) are highlighted.

2.4.2 Gene length is significantly associated with dN/dS values obtained from pairwise and lineage-specific substitutions for *A. thaliana*

dN/dS estimates are also influenced by the fact that pairwise alignments could introduce biases due to branch-specific changes in the strength and direction of selection. For example, if a gene in *A. lyrata* was under a greater degree of purifying selection, this would result in a decreased dN/dS estimate in *A. thaliana* [24]. To address this, we calculated lineage-specific dN/dS values for *A. thaliana* by using *T. parvula* as an outgroup. However, compared to dN/dS estimates based on pairwise alignments with either *A. lyrata* or *T. parvula*, we found the strength of correlations with all sequence characteristics and expression parameters to be reduced (Table S2.1). It is notable that when correlating gene expression level and dN/dS, the estimate of rho is reduced more than 50% using a lineage-specific rather than a pairwise dN/dS. If correlating expression level with

lineage-specific dN/dS, $rho = -0.18$; with pairwise dN/dS estimates, $rho = -0.41$ and -0.29 for alignments against *A. lyrata* and *T. parvula*, respectively (Tables 2.1 and S2.1).

Gene length has been found to be significantly associated with rates of sequence substitution in several species [145, 144, 35, 143]. In *A. thaliana* a significant negative relationship has previously been observed using a sample of 11,492 *A. thaliana – A. lyrata* orthologous pairs [108], although no such relationship is shown in a similar study based on a smaller sample size [63]. Using a similarly broad dataset as that used in ref. [108], our results find a significant association with dN/dS. Given that expression and dN/dS are also negatively correlated, i.e. highly expressed genes are more constrained [63], then it also follows that those genes under stronger purifying selection are more likely to have both a higher expression level and, in plants, possibly a longer length. It is therefore also possible that a relationship between gene length and the rate of protein evolution is a by-product of the relationship between expression and dN/dS. We do not find, however, that the association between dN/dS and gene length is explained by the covariance between gene length and gene expression.

Furthermore, in contrast to the correlation of evolutionary rate with gene expression, the use of lineage-specific dN/dS values for *A. thaliana* did not result in a similar decrease in the correlation of evolutionary rate with gene length. In fact, gene length was found to be the primary predictor of lineage-specific dN/dS. It is possible that the comparatively reduced prominence of expression level as a predictor of evolutionary rate is explained in this case by mating system: *A. thaliana*, unlike *A. lyrata* or *T. parvula*, is a selfing species, and as such may have a reduced efficacy of purifying selection [54]. In this respect, the degree of constraint acting upon highly expressed genes may be partially masked when using lineage-specific dN/dS estimates. Nevertheless, that *A. thaliana* experiences a general trend of relaxed selection compared to *A. lyrata* is only weakly supported [155] and in any case the relationship of expression level to lineage-specific dN/dS for *A. lyrata* is equally reduced, assuming expression to be equivalent in both species (*rho* = -0.15, $p < 2.2 \times 10^{-16}$). In addition, it is important to note that the differences between pairwise and lineage-specific dN/dS are not explained by the differences in gene/codon samples used to estimate dN/dS resulting from the fact that a smaller proportion of the *A. thaliana* genome can be simultaneously aligned with both the *A. lyrata* and the *T. parvula* genomes.

2.4.3 Exon edge removal, but not lineage-specific substitution patterns, unmask higher levels of positive selection

One key objective of this study was to assess whether exon edge conservation and the use of pairwise alignments could be masking additional signatures of molecular adaptation. Our results revealed that the edges of exons are under comparatively higher selective constraint than sequence elsewhere in the gene. By removing these sequences and re-estimating dN/dS for each gene, the distribution of dN/dS values is shifted to the

right, more so than by the random removal of an equivalent amount of sequence. This range is indicative of relaxed purifying, or stronger positive, selection, and accordingly we find that the proportion of genes under potential positive selection ($dN/dS > 1$) is increased. Of particular interest are four genes (AT1G08680, AT1G60930, AT2G17305 and AT4G27370) whereby dN/dS becomes > 1 only after codons are removed sequentially from the exon edges, but not when codons are removed from random positions. This could suggest, in these cases, that an adaptive signature has been partially masked by disproportionate synonymous substitutions at the edges of exons. Of note is that AT1G08680 (ARF GAP-like zinc finger-containing protein ZIGA4) has been linked to adaptive germination phenotypes [156] and that AT1G60930 (RECQ helicase L4B) appears to be a duplicate gene that has undergone a degree of functional divergence [157].

When considering lineage-specific dN/dS, however, the proportion of genes with dN/dS > 1 is significantly lower than when dN/dS is estimated using pairwise alignments of *A. thaliana* with *A. lyrata*. Having found a significant effect of exon edge conservation and lineage-specific substitution upon dN/dS estimates when each was considered separately, we wished to test whether the relationship between dN/dS and the set of genomic parameters changed when both factors are taken into account together. However, there was only a very limited number of genes for which full exons could be aligned across all three species, as required for the analysis of codon removal at the exon edges and the estimates of lineage-specific dN/dS. Using a limited sample ($n = 73$) in which 10 codons could be removed from the exon edges, we found no significant differences in the relationship of dN/dS to any genomic parameter after codons were removed from the exon edges compared to removal at random sites (Tables S2.12 and S2.18). As better annotation of the *A. lyrata* and *T. parvula* genomes, or those of related species, become available it would be possible to assess the effects of exon edge conservation upon dN/dS estimates using lineage-specific substitutions.

In sum, we show that higher conservation at the edges of exons in *A. thaliana* plays an important part in determining dN/dS ratios by increasing the proportion of conserved synonymous sites. The effect of these conserved regions upon overall dN/dS values partly explains the relationship between rates of protein evolution and expression level. By accounting for lineage-specific substitution patterns and the effect of conservation at the exon edges, the ability of expression level to explain variation in evolutionary rate is diminished, with gene length becoming the strongest covariant. In addition, we found evidence of masked positive selection from the conservation of exon edges, irrespective of the noise introduced to dN/dS estimates by the use of pairwise alignments.

2.5 Materials and Methods

2.5.1 Data sources

Exon coordinates for *A. thaliana* strain Col-0 were obtained from The Arabidopsis Information Resource (TAIR) (ftp.arabidopsis.org/, file 'TAIR10_exon_20101028', downloaded 15th February 2013). The *A. lyrata* genome [105], strain MN47 (Entrez genome project ID 41137), was obtained from Genbank (http://www.ncbi.nlm.nih.gov/nuccore/ ADBK00000000, downloaded 17th October 2012). The *T. parvula* genome, version 2.0 [107], was obtained from http://thellungiella.org/blast/db/TpV8-4.fa (downloaded 17th October 2012).

2.5.2 Tests of sequence evolution and selection

Two measures of the degree and direction to which *A. thaliana* sequences diverge from a neutral expectation were calculated – a neutrality index (see below), and dN/dS. Calculations require data both on the number of polymorphic and the number of diverged residues in each sequence. To obtain the former, we used SNP data obtained after aligning 17 fully sequenced and independently assembled accessions against the Col-0 reference genome [39] (data from Po-0 was not used as it has both unusually high heterozygosity and similarity to Oy-0). Diverged positions were identified from pairwise alignments of *A. thaliana* against both *A. lyrata* and *T. parvula*. Alignments were made for 21,198 genes against *A. lyrata* and 10,289 genes against *T. parvula*, of which 7086 genes could be aligned against both. To obtain these alignments, exons corresponding to the longest available transcript per *A. thaliana* gene were matched to each species using the default parameters of blastn [158] and a significance threshold of 1×10^{-10}. Sequence alignments were then obtained using the best hit sequence and the Smith-Waterman algorithm (fasta35 with parameters –a –A) [159]. These alignments were then concatenated to create a single sequence alignment per gene. Finally, to ensure the alignment was in-frame, the translated *A. thaliana* sequence was aligned against either the *A. lyrata* or *T. parvula* sequence using tblastn (default parameters and significance threshold 1×10^{-10}). For genes with at least 150 aligned bases, estimates of dN/dS were calculated from the concatenated sequences using the Yang and Nielson model, as implemented in the yn00 package of PAML [160]. These estimates are referred to as the 'pairwise' dN/dS. We also calculated a lineage-specific estimate of dN/dS using the extremophile crucifer *Thellungiella parvula* [107] as an outgroup, according to the method of [24]. Firstly, we identified those *T. parvula* genes with detectable homology to an *A. thaliana* gene for >50% of the CDS length of the longest Col-0 transcript (blastn, default parameters [158]). Multiple sequence alignments between the CDS of an *A. thaliana* gene, its *A. lyrata* orthologue (if extant) and the homologous sequence in *T. parvula* were then made using PRANK [161]. dN/dS was calculated using the codeml program of PAML [160], with the equilibrium

codon frequencies of the model used as free parameters (CodonFreq = 3). This data was filtered to remove sequences less than 150bp in length or with branches showing either $dS < 0.02$, $dS > 2$ or $dN > 2$ as these are either unreliable for estimates of the dN/dS ratio, non bona fide orthologues or otherwise saturated with substitutions [161]. Finally, to calculate a lineage-specific dN/dS for *A. thaliana*, we assumed an unrooted tree topology of ([*A. thaliana*, *A. lyrata*], *T. parvula*).

The neutrality index for each sequence, NI, was calculated as $log((2D_s+1)(2P_n+1)/(2D_n+1)(2P_s+1))$, where D_n and D_s are the numbers of non-silent and silent substitutions, and P_n and P_s are the numbers of non-silent and silent polymorphisms [162]. NI can be interpreted in the same manner as a McDonald-Kreitman test for comparing the ratio of fixed to within-species differences: its symmetrical distribution allows the inference of purifying selection when NI > 0 and the inference of positive selection when NI < 0 [163].

2.5.3 Exon edge trimming

To assess the effect of purifying selection upon exon edges, we removed up to 30 codons, one at a time, from the edges of each *A. thaliana* exon that could be fully aligned against the *A. lyrata* or *T. parvula* genome with an alignment both in-frame and a multiple of three in length. Exons were then concatenated and genes with sequences of at least 150bp after trimming constituted 'trimmed' subsets of, at minimum, 1443 genes (i.e. those for which all 30 codons can be removed) and 174 genes, respectively. All analyses comparing 'trimmed' and 'untrimmed' sequences use the same set of exons per gene (coordinates given in Tables S2.6, S2.7, S2.8, S2.13, S2.14 and S2.15).

2.5.4 Alternative splicing

Alternative splicing indices were calculated as described in [152]. In brief, alternative splicing events were identified by aligning EST data obtained from dbEST [164] to the genome sequence (ftp://ftp.ncbi.nih.gov/repository/dbEST, downloaded 1st May 2011), as described in [165], and comparing mapping coordinates for each transcript. To correct for the well-known bias in alternative splicing estimates resulting from differential transcript coverage between genes [69, 98, 71], a transcript number normalisation method [98] was implemented as described in [152] whereby the number of alternative splicing events per gene is calculated as the average number of events detected using 100 random samples of 10 ESTs.

2.5.5 Randomisation test

Estimates of dN, dS, dN/dS and NI vary when codons are sequentially removed from the edges of exons, suggesting that the strength of selection differs in these regions. To assess whether this is significant, then for each codon removed at the edge of an exon, we created

a parallel set of estimates of dN, dS, dN/dS and NI after random codon removal (1000 randomisations per gene) for comparison. A numerical p-value was calculated as follows: let q be the number of times the 'sequential removal' estimate of dN, dS, and dN/dS was higher than the 'random removal' estimate (or lower, in the case of NI). Letting $r = s - q$, then the p-value of this test is $(r+1)/(s+1)$. As variable estimates of dN, dS, dN/dS and NI can in turn alter the correlation strength with predictors of evolutionary rate (such as, e.g., expression level), the above test was also repeated using estimates of Spearman's *rho* for both the 'sequential removal' and 'random removal' conditions.

2.5.6 Expression data

Three independent sources of *A. thaliana* transcript abundance data were used:

(1) the Arabidopsis Development Atlas (ADA) generated by the AtGenExpress Consortium [62] (NASCARRAYS reference numbers 150-154, together representing 79 tissues, were obtained from NASC AffyWatch [http://affymetrix.arabidopsis.info/, downloaded 7th November 2011]). Expression level was quantified as both the maximum absolute gcRMA (Robust Multi-array Analysis corrected for the GC-content of the oligo [166]) across all tissue types (clustering the data into seven types – root, stem, leaf, flower, pollen and apex) [63], and as the average across all 79 tissues (with each value itself the mean of triplets) [108]. Breadth of expression was calculated from this database as both the number of tissues in which a gene is expressed and the tissue specificity index (*tau*), a scalar measure bounded between 0 (for housekeeping genes) and 1 (for genes expressed in a single tissue) [149].

(2) Massive parallel signature sequencing (MPSS) data – which quantifies gene expression by counting short (17-20bp) mRNA-derived tags – corresponding to five tissues (http://mpss.udel.edu/at/mpss_index.php, downloaded 28th March 2011) [167, 168, 169]. Expression level was quantified as either the average [170] or the maximum number of tags across all tissues [64].

(3) RNA-seq transcript abundance data, where expression levels were taken as absolute read values corrected by sequence length [39].

On top of the indices of expression obtained from each dataset, all three estimates of transcript abundance (MPSS, ADA and RNA-seq) were transformed into Z-scores [171] to allow direct comparisons between them. In addition, the weighted average of two sets of *A. thaliana* protein abundance data was obtained for a total of 19,761 genes (pax-db.org, downloaded 15th February 2013) [172, 173]. The data employs tandem mass spectrometry to quantify protein abundance by spectral counting.

2.5.7 Other data sources

Codon usage bias per gene was expressed as both the effective number of codons (ENC) [174], and as the frequency of optimal codons (F_{op}) [175]. The number of protein-protein

interactions (PPIs) per gene was obtained from BioGRID, version 3.1.75 [176, 177]. Recombination data was obtained from a previous study [178]; this variable is used as a sanity test as an insignificant relationship between recombination and dN/dS is expected in an effectively obligate selfer. A gene's degree of multifunctionality was considered to be the number of GOslim terms assigned to it for biological processes. 'GOslim' is a condensed set of gene ontology (GO) categories, obtained from TAIR (ftp://ftp.arabidopsis.org/home/tair/Ontologies/Gene_Ontology/

ATH_GO_GOSLIM.txt, downloaded 8th October 2013).

Chapter 3

Presence/absence variation in *A. thaliana* is primarily associated with genomic signatures consistent with relaxed selective constraints

3.1 Summary

The sequencing of multiple genomes of the same plant species has revealed polymorphic gene and exon loss. Genes associated with disease resistance are over-represented amongst those showing structural variations, suggesting an adaptive role for gene and exon presence/absence variation (PAV). To shed light on the possible functional relevance of polymorphic coding region loss and the mechanisms driving this process, we characterised genes which have lost entire exons or their whole coding regions in 17 fully sequenced *Arabidopsis thaliana* accessions. We found that although a significant enrichment in genes associated with certain functional categories is observed, PAV events are largely restricted to genes with signatures of reduced essentiality: PAV genes tend to be newer additions to the genome, tissue specific and lowly expressed. In addition, PAV genes are located in regions of lower gene density and higher transposable element density. Partial coding region PAV events were associated with only a marginal reduction in gene expression level in the affected accession and occurred in genes with higher levels of alternative splicing in the Col-0 accession. Together these results suggest that although adaptive scenarios cannot be ruled out, PAV events can be explained without invoking them.

3.2 Introduction

Intra-species variation in gene content represents an important source of heterogeneity in the genome of a species and potentially contributes to an organism's adaptability in response to external pressures [179]. Cataloguing significant gains and losses in coding regions within or between species will allow a deeper understanding of the mechanisms underlying the molecular evolution of genomes, and can assist in identifying functional variation in agronomically elite varieties of staple crops [180]. To this end, several studies have examined polymorphic full or partial gene loss in several plant species. For instance, after re-sequencing 50 rice genomes, up to 1327 possible gene loss events (2.4% of the total gene set) were detected relative to the Nipponbare reference accession [181]. Significant intra-species variation in gene content has also been reported in maize [65], sorghum [66] and soybean [67]. Previous studies in the model plant *Arabidopsis thaliana*, using re-sequencing microarrays and Illumina sequencing-by-synthesis reads, have also shown significant variations in total nuclear genome sequence among naturally occurring strains [182, 183]. A more recent study using 18 fully sequenced *A. thaliana* genomes found that, relative to the reference accession Col-0, 93.4% of proteins had intra-species variation in their genes, inclusive of large deletions [39] with around 775 genes per accession found to have deletions spanning 50% or more of their coding region sequence [39]. A comparison of 80 *Arabidopsis* genomes found that 9% of the total genes in *A. thaliana* showed presence/absence variation (PAV) averaging 444 absent genes per accession [43].

Characterisation of coding region presence/absence variation has shown certain gene categories to be significantly enriched. For instance, 52 of the 154 nucleotide-binding site leucine-rich repeat (NBS-LRR) R (resistance) genes were found to be deleted in at least one of fifty rice cultivars [181]. Similar over-representation of the R genes in *A. thaliana* has also been observed [184, 185], whilst in the soybean, genes enriched in structural variation are more likely to be involved in nucleotide binding and biotic defence [67]. Enrichment of particular functional gene categories among genes affected by structural polymorphism suggests these structural polymorphisms may have a functional role, allowing accessions to be better adapted to the environmental conditions they face.

However, this hypothesis has not been explicitly tested. If significant polymorphic deletions are adaptive, we would expect that affected genes should show multiple signatures of being under selection. On the other hand, if structural polymorphisms mostly affect genes evolving under relaxed constraints, then their adaptive significance should be questioned. Here we characterise genes affected by presence/absence variation (PAV) spanning whole exons in *A. thaliana*, to investigate which genomic features, if any, are associated with these polymorphisms. Our results provide insights into the likely functional impact of structural variation in protein-coding genes.

3.3 Results

In order to characterise presence/absence variation (PAV) in *A. thaliana*, we examined previously identified polymorphic deletions in 17 fully sequenced *Arabidopsis* accessions for which transcriptome data was available [39] (see Materials and Methods). We compiled a set of deletions that spanned entire exons in any of 17 accessions relative to the Col-0 reference genome. A subset of the annotated deletions was experimentally validated [39]. To further rule out the possibility of wrongly identifying deletions due to differences between assemblies, exons were confirmed as missing by searching for homology between the Col-0 exon on all other accessions (see Materials and Methods). A total of 794 exons were classified as missing in at least one of 17 accessions, corresponding to 411 genes (approx. 1.5% of the total gene set) including 81 genes where the full coding region was completely absent in at least one accession (Table S3.1). Exon losses are not uniformly distributed throughout the gene: missing exon sequences are more often found nearer the ends of each gene (Figure 3.1).

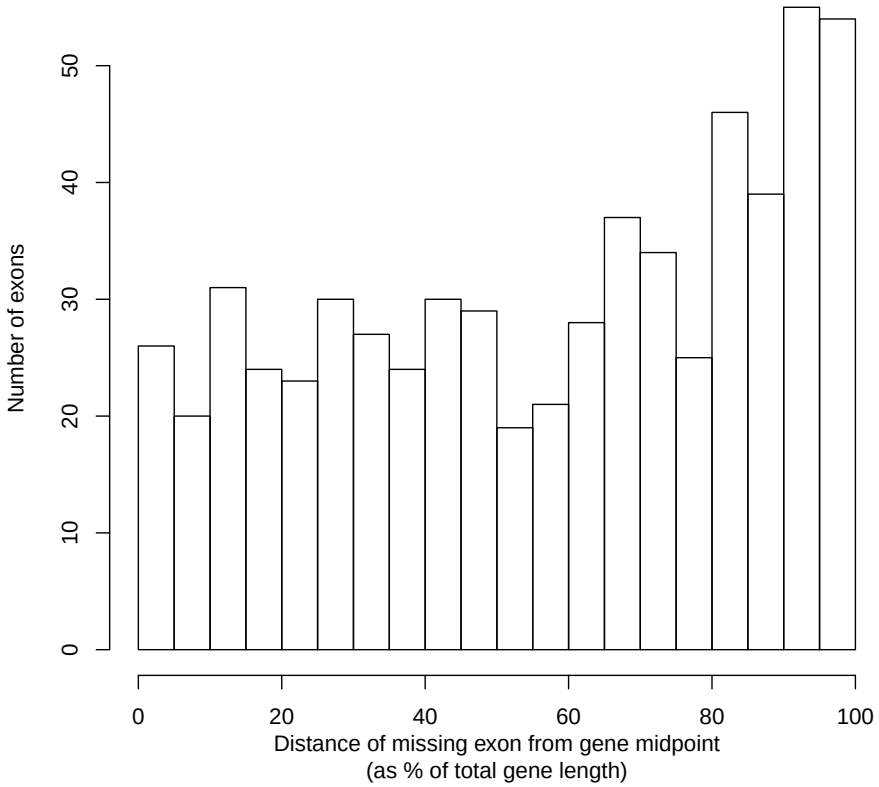

Figure 3.1: Location within a Col-0 gene of exons missing in at least one other accession.

Only genes with at least one, but not all, exons missing were considered. For the regression of histogram density estimates against bin midpoints, adj. $r^2 = 0.47$, $p = 5 \times 10^{-4}$.

Overall, approx. 0.3% of the genes in each accession have at least one missing exon,

representing between 10-50kb of missing sequence per accession (Table S3.2). A total of 200 genes had exon loss affecting more than one accession, consistent with a previous study reporting a 'common history' to deletion events in *A. thaliana* [186].

Because partial deletions spanning whole exons might have distinct functional implications compared to full coding region deletions, the 330 genes with partial coding region loss spanning at least one full exon in at least one accession (exon presence/absence variation, E-PAV) and the 81 genes with full coding region polymorphic deletions affecting at least one accession (full coding DNA sequence presence/absence variation, CDS-PAV) were examined separately.

3.3.1 Genes involved in signal transduction and both nucleotide and protein binding are over-represented among PAV genes

In order to characterise PAV genes, we first assessed whether these genes were over-represented in particular gene classes or gene ontology (GO) categories. To do so, we used four classification schemes: 'GO', a condensed set of GO terms ('GOslim'), the Pfam protein domain database and the family classification scheme of [39] (see Materials and Methods). Of the 330 E-PAV genes we found most to be poorly characterised with 50% of them having no associated GOslim term. The proportion of poorly characterised genes is greater among CDS-PAV genes, with more than 60% having no associated GOslim term for biological process. When examining genes with associated GOslim terms we found both E-PAV and CDS-PAV genes to be significantly enriched in genes associated with signal transduction and nucleotide binding (Figure 3.2 and Figure 3.3). Furthermore, E-PAV genes also appear significantly enriched in genes associated with the GOslim term 'other binding', which includes proteins that bind to lipids, metal ions and ATP, among other cofactors (Figure 3.2). Significant over-representation of functional categories among PAV genes is consistent with a previous assessment of large coding region indels in the soybean genome [67] and of whole gene deletions in *A. thaliana* [43]. This is also observed when classifying genes using the broader set of 'GO', rather than 'GOslim' terms (Figure 3.4).

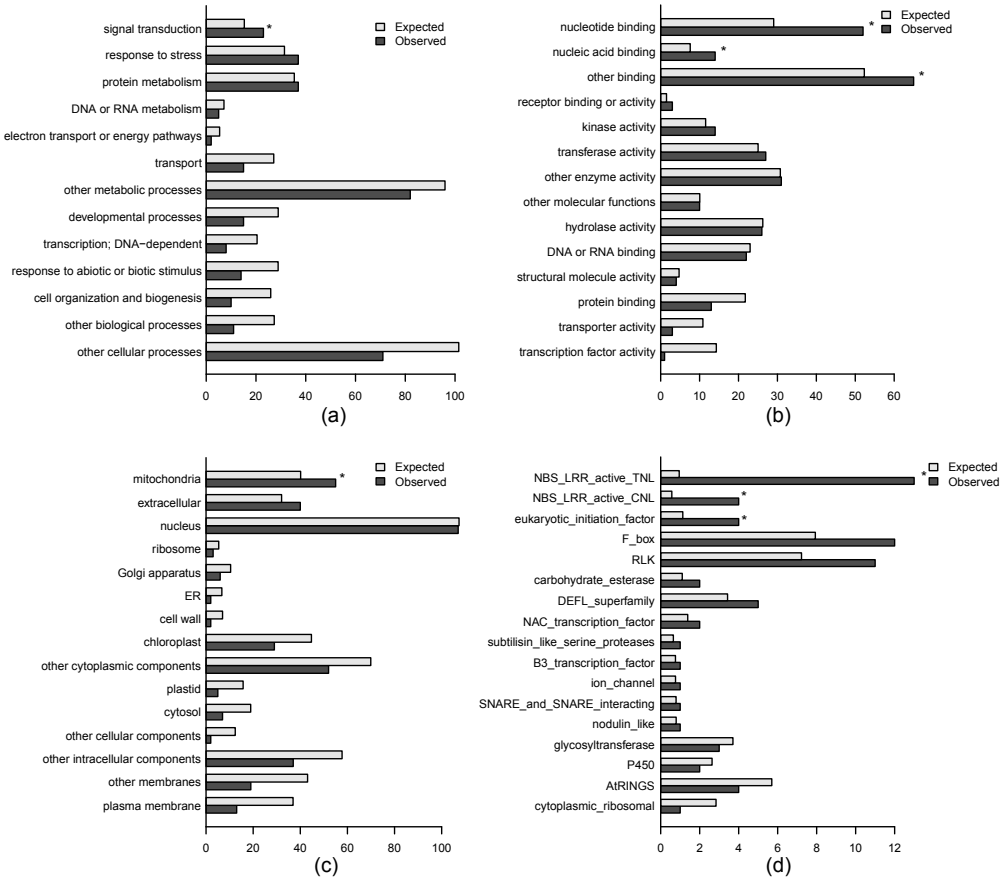

Figure 3.2: Distribution of 'exon presence/absence' (E-PAV) genes ($n = 330$) – those with at least one, but not all, exons missing in at least one accession – by GOslim categories for molecular function (a), biological process (b) and cellular component (c), and by family (d).
Both expected and observed number of E-PAV genes per category represented on each bar. Where there is a significant enrichment ($p <= 0.05$) between the amount of observed and expected E-PAV genes for a particular category an asterisk is shown over the bars. Only categories with at least one E-PAV gene are shown.

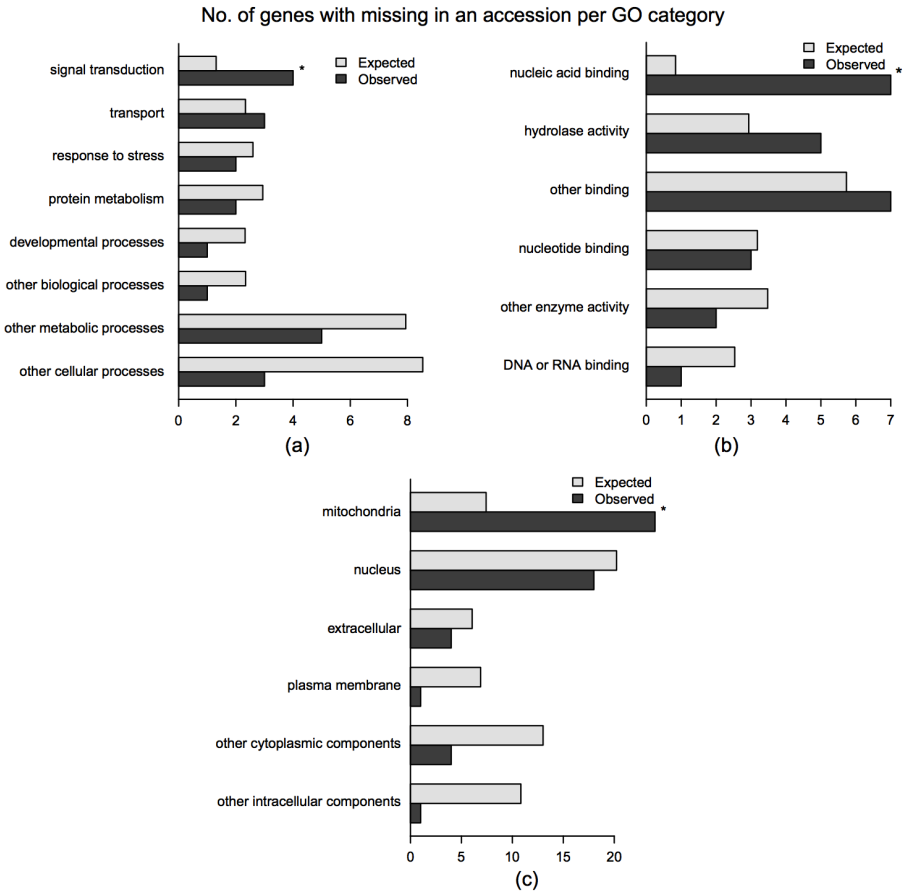

Figure 3.3: Distribution of 'CDS presence/absence' (CDS-PAV) genes ($n = 81$) – those with their entire coding region missing in at least one accession – by GOslim categories for molecular function (a), biological process (b) and cellular component (c).
Both expected and observed number of CDS-PAV affected genes per category represented on each bar. Where there is a significant enrichment ($p <= 0.05$) between the amount of observed and expected CDS-PAV genes for a particular category an asterisk is shown over the bars.

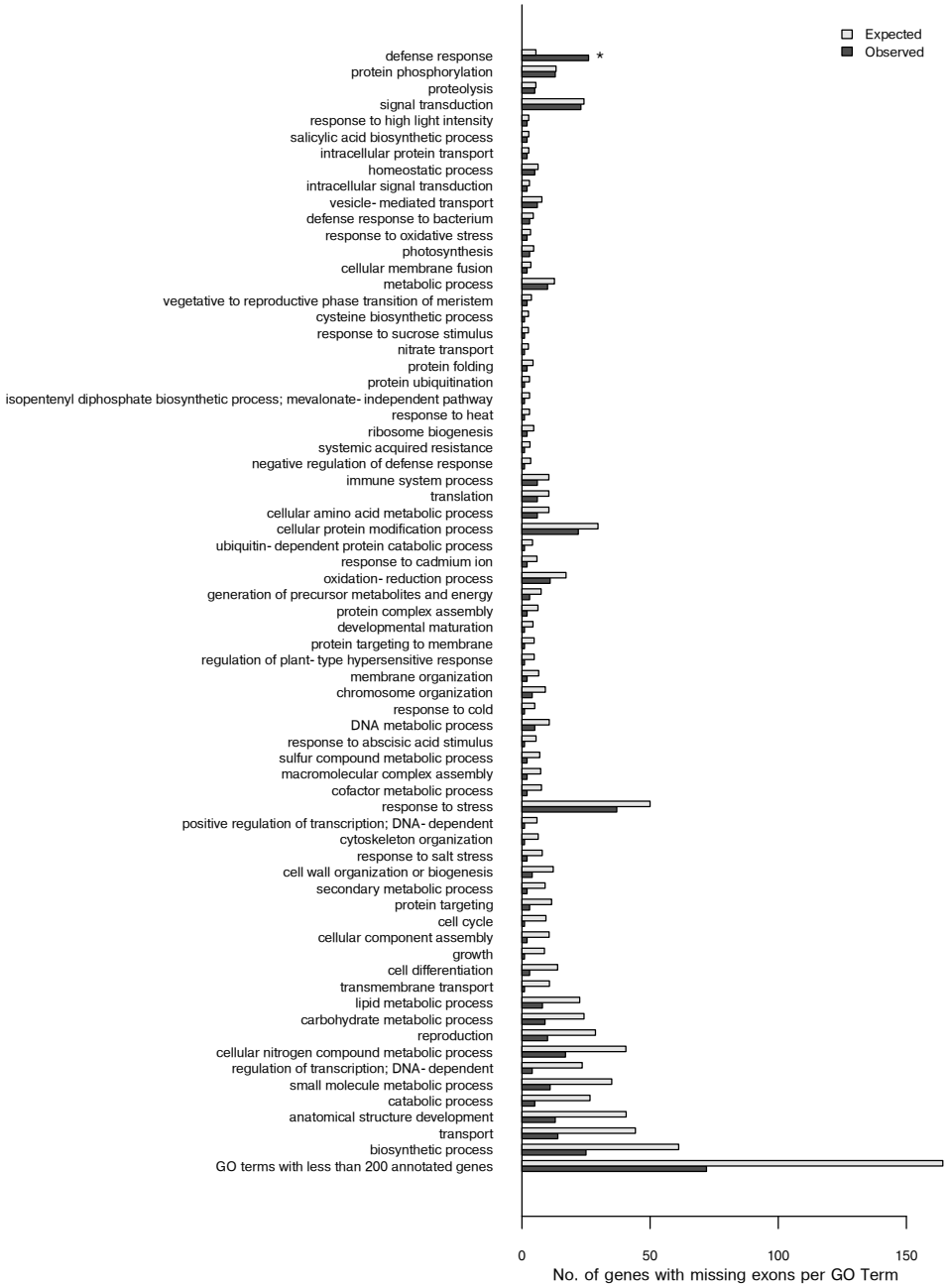

Figure 3.4: Distribution of 'exon presence/absence' (E-PAV) genes – those with at least one, but not all, exons missing in at least one accession – by GO category ($n = 330$). Only GO terms with 200 or greater genes were considered. Both expected and observed number of E-PAV genes per category represented on each bar. Where there is a significant enrichment ($p <= 0.05$) between the amount of observed and expected E-PAV genes for a particular category an asterisk is shown over the bars.

When classifying genes by family we observe an over-representation of members of the NBS-LRR (nucleotide binding site leucine rich repeat) family – involved in pathogen detection [187] – among E-PAV genes (families 'NBS-LRR active TNL', adjusted p-value = 8.57×10^{-35}, and 'NBS-LRR active CNL', adjusted p-value = 4.63×10^{-5}; Figure 3.2),

consistent with previous findings [185]. Furthermore, when examining the 3753 Pfam ID gene associations (Figures 3.5 and 3.6) we observe an over-representation of members of the NB-ARC and LRR domain containing families (note that 'NBS-LRR' refers to a composite of the NBS and LRR domains and that the NBS domain is also known as 'NB-ARC' [68]). No enrichment of any particular gene family was observed among CDS-PAV genes (data not shown).

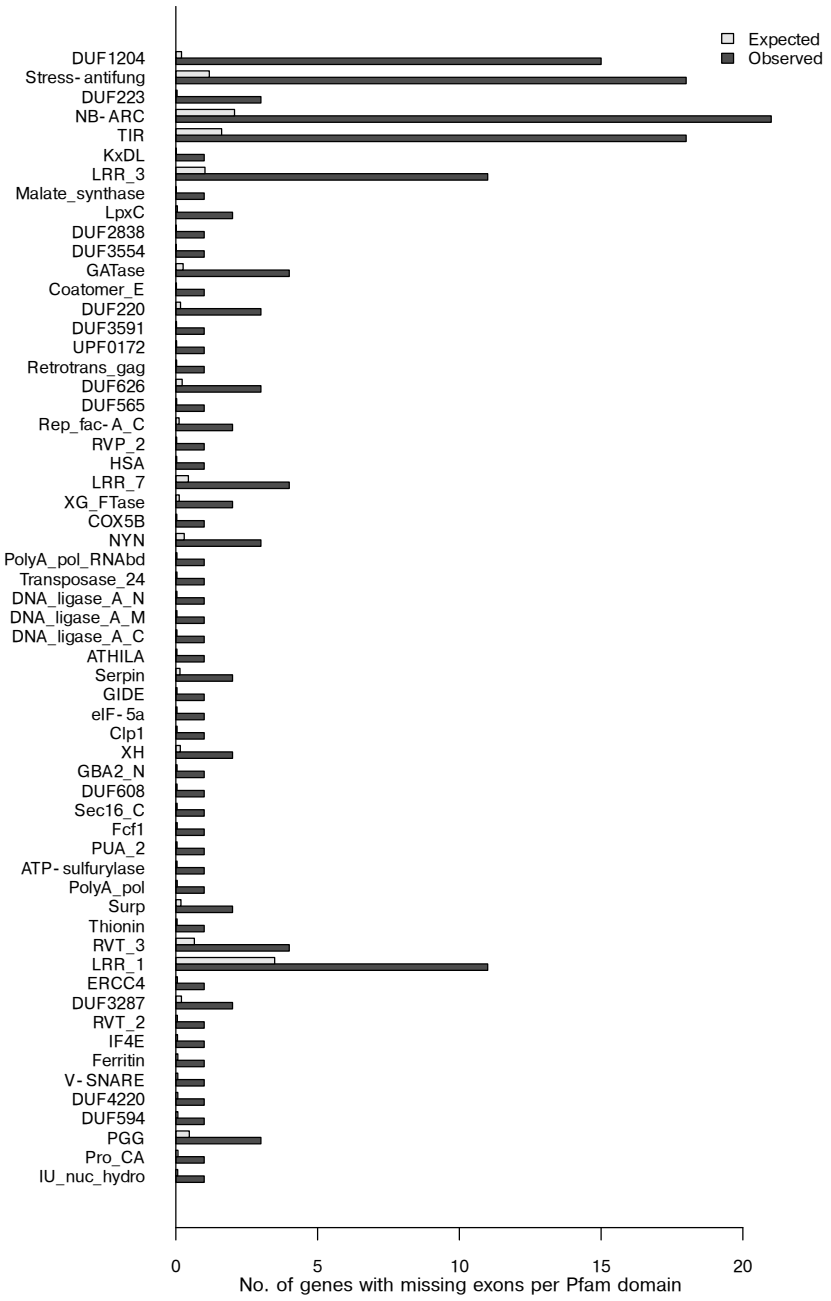

Figure 3.5: Distribution of 'exon presence/absence' (E-PAV) genes – those with at least one, but not all, exons missing in at least one accession – by Pfam category ($n = 330$). Both expected and observed number of E-PAV genes per category represented on each bar. Only categories with a significant enrichment are shown. Note that there is a significant enrichment of E-PAV genes in the category of 'no Pfam annotation' (adjusted p-value = 3.96×10^{-5}).

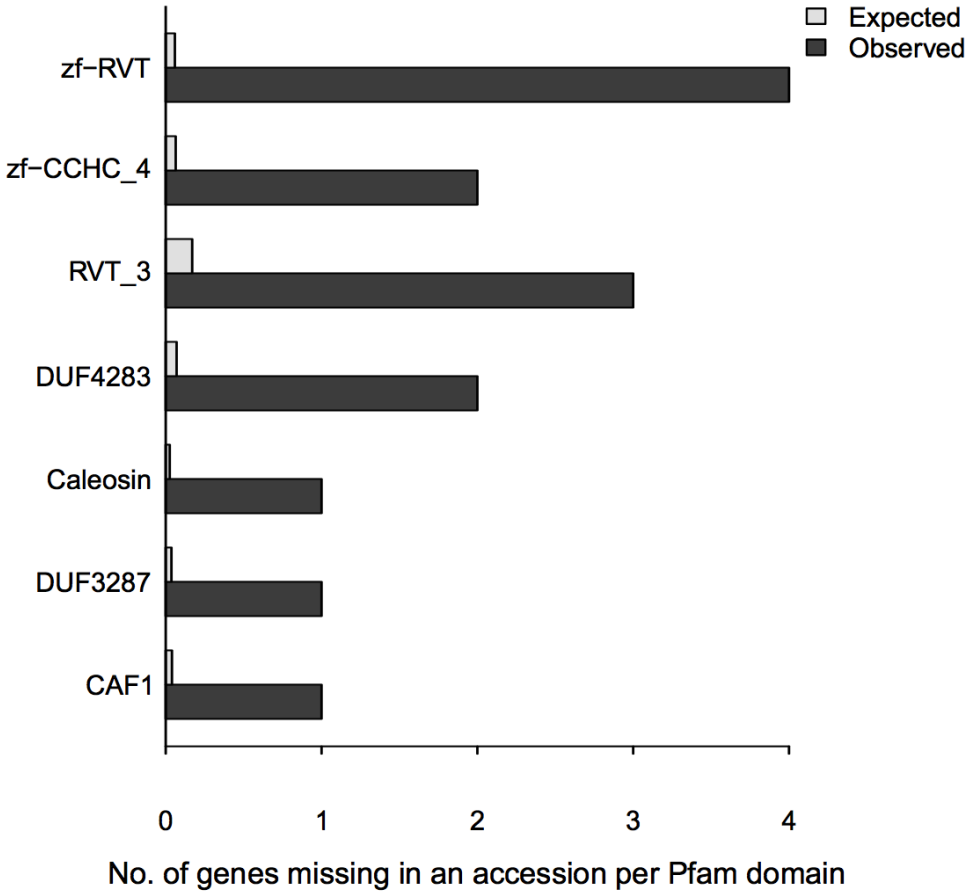

Figure 3.6: Distribution of 'CDS presence/absence' (CDS-PAV) genes – those with their entire coding region missing in at least one accession – by Pfam category ($n = 81$). Both expected and observed number of CDS-PAV genes per category represented on each bar. Only categories with a significant enrichment are shown.

These significant enrichments in gene functional and domain annotations are in line with previous findings in *Arabidopsis* [43] and other plant species [67, 65, 66] and have been proposed to reflect the adaptive role of large polymorphic deletions.

3.3.2 Genes affected by PAV show signatures consistent with relaxed selective constraints

To determine if PAV genes are generally associated with fast evolving proteins potentially under positive selection, we examined the rates of non-synonymous to synonymous changes per gene (dN/dS). Using a randomisation test, E-PAV genes were found to have a significantly higher dN/dS ratio compared to genes with all exons present but only 8 genes have a dN/dS ratio above 1 (Figure 3.7, Table 3.1 and Table S3.1). CDS-PAV genes had a non-significant increase in dN/dS compared to intact genes (those not affected by deletions spanning at least one exon in any accession; Figure 3.7 and Table 3.1).

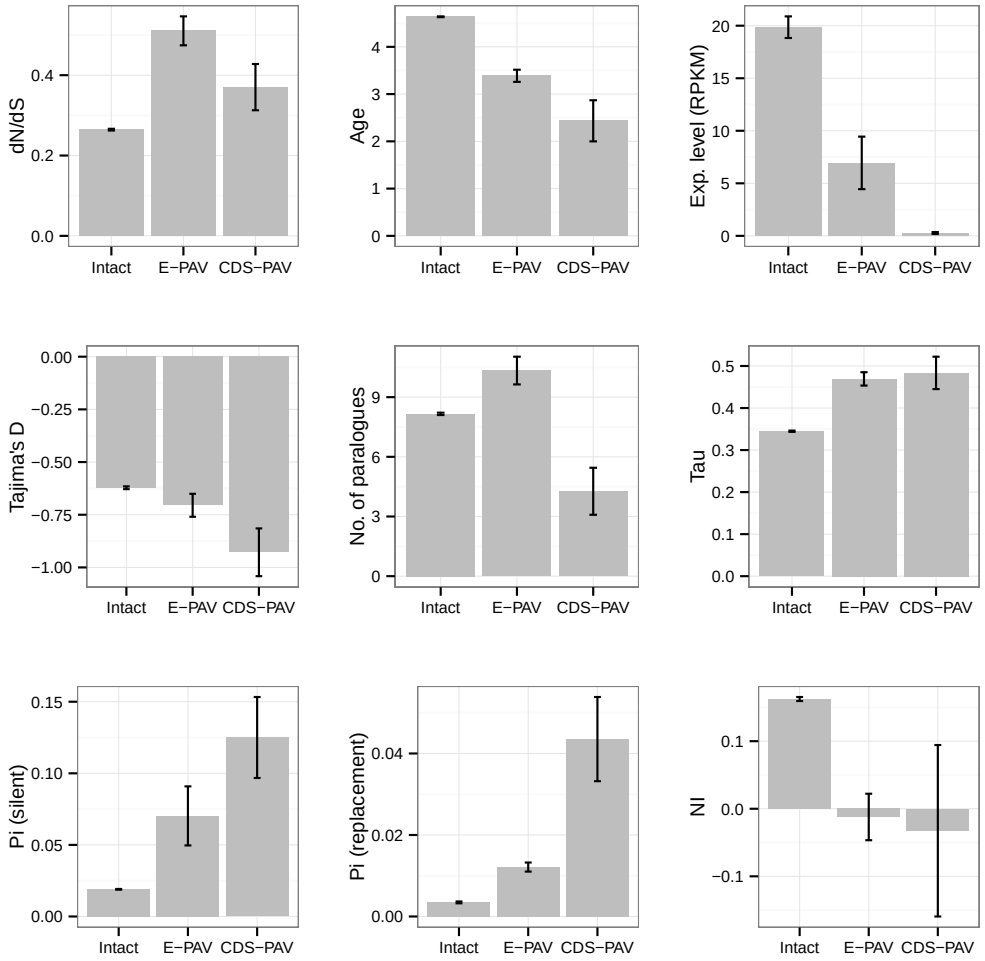

Figure 3.7: Genetic features associated with intact (having no exons under P/A variation), E-PAV (having at least one, but not all, exons missing in at least one accession), and CDS-PAV (having the entire CDS missing in at least one accession) genes.
From left to right, top to bottom: dN/dS, age, expression level, Tajima's D, number of paralogues and *tau*. See Table 3.1 for values of means and statistical analysis.

Characteristic (average)	Intact genes	E-PAV genes		CDS-PAV genes	
		Value	p	Value	p
dN/dS	0.26	0.51	$1.00\text{x}10^{-4}$	0.37	0.07
NI	0.16	-0.01	$1.00\text{x}10^{-4}$	-0.03	0.05
Tajima's *D*	-0.62	-0.71	0.09	-0.93	$7.60\text{x}10^{-3}$
Age category	4.64	3.39	$1.00\text{x}10^{-4}$	2.43	$1.00\text{x}10^{-4}$
Number of paralogues	8.16	10.34	$1.00\text{x}10^{-4}$	4.27	$1.00\text{x}10^{-4}$
Number of alt. splicing events	1.13	3.32	0.05	2.2	0.05
Expression level (RPKM)	19.86	6.94	$2.00\text{x}10^{-4}$	0.27	$1.00\text{x}10^{-4}$
Tau	0.34	0.47	$1.00\text{x}10^{-4}$	0.48	0
Exon length (bp)	284.22	170.08	$1.00\text{x}10^{-4}$	312.8	0.17
Gene length (bp)	2142.42	2360.49	$8.30\text{x}10^{-3}$	639.91	$1.00\text{x}10^{-4}$
Intergenic distance (bp)	2189.76	3548.47	$1.30\text{x}10^{-3}$	5605.37	$1.80\text{x}10^{-3}$
Distance to nearest TE (any) (bp)	5729.75	2507.17	$1.00\text{x}10^{-4}$	1564.68	$1.00\text{x}10^{-4}$
Distance to nearest TE (DNA) (bp)	11359.44	5940.86	$1.00\text{x}10^{-4}$	4906.35	$1.00\text{x}10^{-4}$
Distance to nearest TE (LINE) (bp)	58179.69	30359.92	$1.00\text{x}10^{-4}$	20190.59	$1.00\text{x}10^{-4}$
Distance to nearest TE (RC) (bp)	46743.44	24944.49	$1.00\text{x}10^{-4}$	16747.54	$1.00\text{x}10^{-4}$
Distance to nearest TE (SINE) (bp)	338168.3	175986.8	$1.00\text{x}10^{-4}$	129272.1	$1.00\text{x}10^{-4}$
Distance to nearest TE (LTR) (bp)	293305.8	153643.7	$1.00\text{x}10^{-4}$	111584	$1.00\text{x}10^{-4}$
Nucleotide diversity (silent)	0.02	0.07	$1.00\text{x}10^{-4}$	0.13	$1.00\text{x}10^{-4}$
Nucleotide diversity (replacement)	$3.45\text{x}10^{-3}$	0.01	0.01	0.04	$3.70\text{x}10^{-3}$

Table 3.1: Characteristics of E-PAV and CDS-PAV genes compared to genes with all exons present in all accessions.

To further examine the selective pressures associated with PAV genes, we considered nucleotide diversity at both replacement sites and silent sites (defined as non-coding sites and the synonymous sites of protein-coding regions) for each gene, according to [39], and calculated a neutrality index (NI) to compare the ratio of fixed to within-species differences. PAV genes were found to be associated with higher nucleotide diversity in both silent and replacement sites (Table 3.1) and to have, on average, a lower NI. While taken together, a lower NI, a higher dN/dS ratio and a higher nucleotide diversity are suggestive of relaxed selective constraints this pattern is also consistent with a scenario of positive and/or balancing selection. To differentiate between these possible scenarios, Tajima's D was calculated for each gene (see Materials and Methods). A threshold of ± 2 was considered as the point at which D significantly departs from the null expectation of neutral evolution for any given gene. Of the 330 E-PAV genes with exon presence/absence variation, 24 have $D < -2$ and only 2 have $D > 2$ (AT1G12180, $D = 2.17$, and AT5G35460, $D = 2.05$, both of which are functionally uncharacterized). Among CDS-PAV genes, only 7 have $D < -2$ and none have $D > 2$. Compared to the set of intact genes, there are no significant differences in the proportion of PAV genes either with $D < 2$ (randomisation test $p = 1$ for both E- and CDS-PAV genes) or $D > 2$ (randomisation test $p = 0.93$ and $p = 1$ for E- and CDS-PAV genes, respectively). As demographic characteristics of the *Arabidopsis* population may result in a shift in the average Tajima's D among the general pool of genes it is possible that these hard thresholds may not be informative. Indeed, we find that intact genes in *Arabidopsis* have the average Tajima's D estimate shifted towards

negative values. Thus, PAV genes could fall short of the hard threshold of $+2$ and still have a higher D estimate than the general pool of genes, suggestive of balancing selection. However, E-PAV genes do not show significant differences in Tajima's D estimates compared to intact genes and CDS-PAV genes have, in fact, a significantly lower estimate of D (Figure 3.7, Tables 3.1 and S3.1). It is possible that PAV genes may have a higher range of D values compared to intact genes, hiding a higher proportion of genes under positive and balancing selection which would not be reflected in overall changes in the mean. To test this, we compared the distributions of Tajima's D estimates in the three sets of genes (intact, E-PAV and CDS-PAV). However, we did not observe any evidence for increased dispersion in D among PAV genes (Figure 3.8). To further examine this possibility, we examined the proportion of PAV genes below the fifth and above the 95th percentile of the 'intact' distribution ($D = -2.05$ and 1.39, respectively). If a significantly higher proportion of E-PAV or CDS-PAV genes are found compared to the intact set at the positive end of the distribution, we can infer the existence of a detectable subset of PAV genes that may be undergoing balancing selection. However, this is clearly not seen – only 2.33% of E-PAV, and no CDS-PAV genes, exceed the threshold value. At the opposite end of the distribution, we observe no overrepresentation in the proportion of E-PAV genes whose estimates of D are lower than the threshold (3.32%) although do observe this for CDS-PAV genes (8.82%). This finding would suggest that a significant proportion of CDS-PAV genes might be undergoing stronger purifying or positive selection relative to intact genes. Together these results suggest that while we cannot rule out the effect of balancing selection acting on a few individual PAV genes a general trend of balancing selection for PAV genes does not readily apply. Indeed, given *A. thaliana* is a near obligate selfer, selection is expected to be less efficacious than in an outcrossing lineage [51], consistent with a general trend of relaxed selection observed against weakly deleterious mutations in this species [54]. It is intrinsically unlikely that in the absence of overdominance, heterozygotes will remain unfixed in successive generations of the population. The excess of negative D values among PAV genes coupled with the higher levels of nucleotide diversity and the significant increases in dN/dS ratios are consistent with a scenario of weaker purifying selection but could also be explained by positive selection.

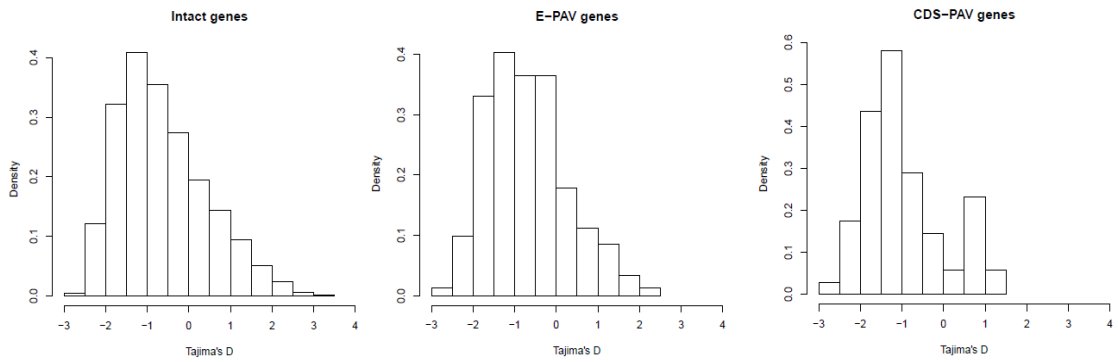

Figure 3.8: Distribution of Tajima's *D* values for intact (having no exons under P/A variation), E-PAV (having at least one, but not all, exons missing in at least one accession), and CDS-PAV (having the entire CDS missing in at least one accession) genes.

We examined a number of parameters which have been previously associated by some studies with gene essentiality to further explore the functional importance of PAV genes, including a gene's age [18] and the number of paralogues it has [188, 189], along with weaker associations such as expression level [112] and tissue specificity [190].

Compared to newer genes, older genes are more likely to be essential [18]. After using the phylogenetic relationships of plant genomes to create a proxy for gene age, we observed that the 330 genes affected by E-PAV are more likely to be newer additions to the genome (Figure 3.7 and Table 3.1). It is also possible that E-PAV genes have a greater number of paralogous genes which might compensate for any loss of function. Consistent with this, we find that those genes with missing exons have a higher number of paralogues compared to those genes with all exons present (Figure 3.7 and Table 3.1). However, the opposite result was observed when analysing CDS-PAV genes – these have an average of 4.2 paralogues compared to genes with no exon losses (Figure 3.7 and Table 3.1), suggesting their function is less essential. This is also consistent with the younger age of CDS-PAV genes. We then assessed the expression patterns of genes affected by exon presence/absence, since broadly and highly expressed genes are typically associated with higher levels of selection [170]. Using a randomisation test, we found that genes with exon losses in one or more accessions, when compared to intact genes, had lower expression levels and higher tissue specificity (Table 3.1). In addition we also observed that exons missing in at least one accession are, on average, shorter than exons present in all accessions (170bp *vs.* 284bp, randomisation test $p = 9.9 \times 10^{-5}$; Table 3.1). However, although exons affected by polymorphic deletions are shorter on average compared to non-deleted exons, E-PAV genes are longer than unaffected genes (2360bp compared to 2142bp, respectively; randomisation test $p = 0.008$; Table 3.1). By contrast, CDS-PAV genes – where polymorphic deletions encompass the gene's entire coding region – were found to be shorter than unaffected genes (640bp compared to 2142bp, randomisation test $p = 9.9 \times 10^{-5}$; Table 3.1).

Overall, these findings show that although certain functional categories are over-represented among genes with exon loss, more generally significant coding region loss is prevalent

amongst novel, lowly expressed and poorly functionally characterised genes. These genes seem to have evolved more recently in the *Arabidopsis* genome and are likely to be under reduced selective constraint.

3.3.3 PAV genes are located in genomic regions that are gene-poor and transposable element-rich

When characterising the genomic context of genes affected by PAV we found that genes with both exon and full coding region loss are separated by longer intergenic distances (Figure 3.9 and Table 3.1). Transposable element density around PAV genes was then assessed as gene-poor areas have been associated with a higher transposable element density [191]. To do this, we used the reference accession (Col-0) and calculated TE density for each gene in all intergenic sequence in 1-100kb windows centred on each gene's midpoint, by counting the number of bases found within TE annotations (see Materials and Methods). E-PAV genes were found to have an approximately two-fold increase in the amount of bases annotated as a TE compared to genes which are intact in all accessions (e.g. TE sequence accounts for approx. 30% of the non-genic sequence within a 10kb window surrounding an E-PAV gene; Figure 3.9 and Table S3.3). Significant enrichment of specific transposable element superfamilies was also observed, notably DNA transposons and LTR retrotransposons (Table S3.3).

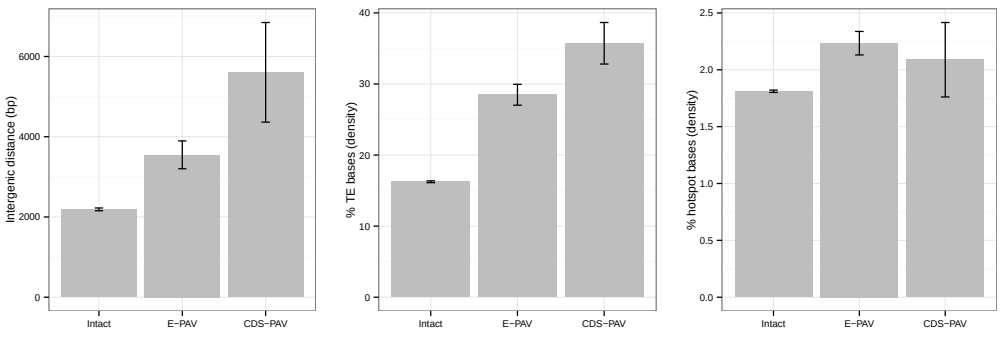

Figure 3.9: Genomic context for intact (having no exons under P/A variation), E-PAV (having at least one, but not all, exons missing in at least one accession), and CDS-PAV (having the entire CDS missing in at least one accession) genes.
Averaged values for the genes in each set are given for, from left to right, the intergenic distance, the percentage of TE bases in the non-genic sequence of a 10kb window centred on that gene's midpoint, and the percentage of recombinogenic motifs in the genic sequence of a 1kb window centred on that gene's midpoint. See also Tables S3.2 and S3.3 for the values of specific TE families and other window sizes.

In addition, we found that genes with missing exons have, on average, a shorter distance from the gene boundary to the nearest TE than those genes with all exons present (2.5kb compared to 5.7kb; randomisation test, $p = 9.9 \times 10^{-5}$; Table 3.1). If calculating the minimum distance to the nearest TE, classified by superfamily, E-PAV genes are significantly

closer to every TE type: rolling circle TEs, DNA transposons, LTR retrotransposons, LINEs and SINEs (Table 3.1). Similar findings were obtained when analysing TE content in the surrounding regions of CDS-PAV genes (Table 3.1).

Certain TE sequence motifs have been associated with recombination hotspots which could drive exon loss through promoting ectopic recombination events [192, 193]. To explore whether genes affected by PAV have a local enrichment for such hotspot motifs, we examined the density of these motifs both in and around genes (see Materials and Methods). However, we observed no significant differences in hotspot motif occupancy in the non-genic regions of windows surrounding E-PAV genes compared to intact genes (in window sizes of 1kb to 100kb centred on the gene's midpoint; Table S3.3). Nevertheless, a significant enrichment in hotspot motif occupancy was observed in the genic sequence of all windows centred on E-PAV genes compared to those centred on 'intact' genes (Figure 3.9 and Table S3.3). When comparing CDS-PAV genes to the intact set, we observed no consistent pattern of higher hotspot motif density within genic regions and only a marginally higher proportion of hotspot motifs in the non-genic regions that surround them, in windows up to 3kb in size ($p < 0.01$; Table S3.4). Taken together, these results show that PAV genes are located in gene-poor and TE-rich regions of the genome further supporting the hypothesis that PAV is associated with relaxed selective constraints. Enrichments of sequence motifs previously associated with recombination hotspots in or around PAV genes suggest that at least some exon deletion events may have resulted from recombination events involving these recombination hotspot motifs.

3.3.4 Exon loss is associated with a marginal reduction in expression level

The above results suggest that exon presence/absence variation is associated with reduced selective constraints. To assess whether exon loss is likely to have resulted in reduced functionality for the genes affected, we compared expression levels for genes with and without missing exons across accessions. If exon loss causes or follows from diminished functionality by previous mutations we would expect, given sufficient time, expression to be significantly reduced in those accessions affected by E-PAV. Using RNA-seq transcription profiles for each *Arabidopsis* accession [39], we compared the expression patterns of individual genes in accessions affected by exon deletions with those accessions where the gene remained intact. To do this, we transformed expression data per accession to Z-scores [171]. We then looked only at those genes where exon loss had occurred in a single accession (210 genes). For each gene, we took (a) the expression level of that gene in the affected accession, and (b) the mean expression level of that gene across the 17 unaffected accessions (the other 16 under study plus the reference genome, Col-0). We found that half of the genes examined had an expression level below this mean and 37% an expression level equal to it. However, on average, expression levels in the affected accession departed little from mean expression in unaffected accessions (0.15 standard deviations).

In 27 genes (13% of cases), expression level in a gene affected by an exon deletion was higher than the mean expression across unaffected accessions with 14 cases showing a statistically significant difference (Figure 3.10 and Table S3.5). These 27 genes are generally poorly characterised with 12 having no functional category annotations. Most genes affected by exon deletions had low expression levels to begin with, although some exceptions are notable, such as rotamase CYP4 (AT3G62030; involved in a variety of cellular functions related to metabolism and response to several types of stress), which has an average expression level in the unaffected accessions of 400rpkm, among the top 1% of genes with detectable expression in Col-0.

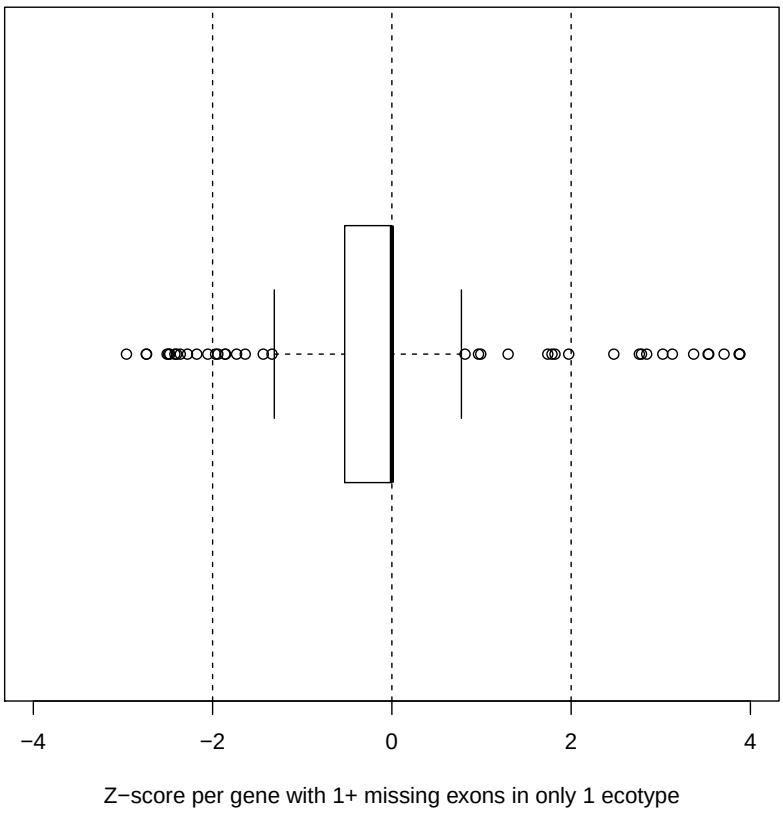

Figure 3.10: Distribution of Z-scores for standardised transcript abundance data in the affected accession.
Data shows 210 genes that have one or more missing exons in only one of 17 *A. thaliana* accessions (relative to Col-0).

It is possible that the moderate effect of exon loss on gene expression levels is explained by an over-representation of alternatively spliced exons amongst the set of missing exons. This would allow for the production of viable protein products in their absence. In order to test this, we quantified alternative splicing in 15,540 *Arabidopsis* genes including 103 of the 330 E-PAV associated genes using a 'comparable alternative splicing index' (see Materials and Methods) which corrects for the distorting effect of variation in transcript coverage among genes (reviewed in [165]). E-PAV genes were found to have a signif-

icantly higher number of alternative splicing events compared to intact genes (3.35 and 1.13 respectively; randomisation test $p = 0.046$).

Overall, these findings suggest that exon losses have only a marginal effect on the expression profile of genes in the accessions affected. The higher levels of alternative splicing among genes affected by exon loss raises the possibility that a significant proportion of lost exons are normally alternatively spliced, reducing selection pressure on these exons since a functional protein product would be produced in their absence anyway.

3.4 Discussion

Intra-species structural variations in genes have been proposed to play an important role in the adaptation of particular populations to variation in environmental conditions [179]. Here we have characterised presence/absence coding sequence variation (PAV) in 17 fully sequenced *A. thaliana* genomes, relative to the reference accession Col-0, affecting 411 genes including 81 instances of whole coding region deletions. We found a significant enrichment of genes associated with the GO terms for protein and nucleotide binding as well as signal transduction. Both gene family and Pfam annotation enrichment analysis revealed significant enrichments of gene members from the disease resistance associated NBS-LRR gene families. Significant deviations from random expectations have been observed in previous studies of PAV genes in plants, with similar over-representation of resistance-associated gene families among PAV genes. For instance, in sorghum [66], PAV genes are enriched in nine Pfam categories, including the NB-ARC domain-containing family. In soybean [67], PAV-affected genes have also been found to be enriched for members of the NB-ARC family, and within the GO category of 'defence response.' CDS-PAV genes have also been shown to deviate from random expectations in *Arabidopsis* [43], with the greatest significant enrichment in PAV genes also reported for those with NB-ARC domains.

These functional and/or gene family enrichments can be suggestive of an adaptive role for PAV events by aiding specific ecotypes in adapting to their local environment. Our results – showing that genes associated with, e.g., resistance are more likely to be affected by PAV – are, at first glance, consistent with this hypothesis. In addition, we were able to confirm a previous report of CDS-PAV for three members of the R gene family – the single-exon gene AT5G05400, and the multi-exon genes AT5G18350 and AT5G49140 [185] – a family known to have signatures of positive selection in *A. thaliana* [194]. However, comprehensive analysis for evidence of selection does not support this as a general interpretation.

dN/dS ratios are one of the most widely used estimates of selective pressure acting on protein coding genes with $dN/dS >> 1$ indicative but not a definitive signature of positive selection [102]. dN/dS estimates typically assume a homogeneity of substitution pattern between lineages although this does not always hold [195]. Nevertheless, *A. lyrata* has

an excess of low-frequency non-synonymous polymorphisms both within and between populations, more restricted in their population distribution but otherwise similar to *A. thaliana* and in both cases consistent with a pattern of weak purifying selection [64]. Although there are, on average, a higher number of substitutions in E-PAV genes compared to intact genes, this is not a clear signature of adaptation, and can suggest comparatively relaxed negative, rather than stronger positive, selection.

We further found that PAV genes have significantly higher nucleotide diversity both at silent and replacement sites, and a lower NI. Nevertheless, *A. thaliana* has a structured population with strong local adaptation to climactic variation [196]. Its excess of non-synonymous polymorphisms - which would bias NI estimates upwards - and reported lack of global sweep signals can be interpreted as low levels of adaptation but are also both expected under a model of predominantly local adaptation [197]. As such, although these observations are suggestive of weaker purifying selection, they can also be expected if PAV genes were under higher balancing selection. Indeed, there is evidence to suggest that the diversity of resistance-associated genes is maintained by balancing selection [198], which are over-represented among PAV genes. Balancing selection has been proposed to stably maintain both the intact gene and the absent allele [43].

So, is balancing selection the most parsimonious explanation for why PAV genes are associated with higher nucleotide diversity? A classic scenario of trans-species polymorphism, associated with balancing selection, cannot be assessed given the limited sequence variation data available for *A. lyrata*, *A. thaliana*'s closest sequenced relative. It is possible that the 'gene/exon present' and the 'gene /exon absent' alleles are under selection to be maintained in different *A. thaliana* populations, allowing them to better adapt to their local environment. This would be consistent with the increase in nucleotide diversity but this scenario cannot be distinguished from alternative neutral models. Conditional neutrality at PAV loci, where the functional gene has ceased to be adaptive in some but not all environments, cannot be ruled out (e.g. in the case of resistance genes where the corresponding pathogen is absent [199]). In this case, the absent allele would have no selective advantage at any point but rather result from relaxed constraints associated with PAV genes in some *Arabidopsis* populations. Moreover, a model of generalised relaxed constraints affecting the PAV loci would also lead to increased nucleotide diversity and slight increases in dN/dS.

Tajima's D, a comparison of two estimators of ϑ (the population mutation rate $4N_e\mu$) – the number of segregating sites and the average number of pairwise differences between sequences [200] – offers a more reliable estimate of selective pressures acting on a gene as it incorporates information about the distribution of segregating alleles in a species. This allows more accurate estimations of the degree and direction of departure of sequence evolution from a neutral expectation (although non-selectionist interpretations of D are also possible, such as recent population expansion or bottlenecking for negative and positive D, respectively) [200]. Tajima's D values do not provide evidence for either E-PAV

or CDS-PAV genes to be under balancing selection. Taking dN/dS, NI, nucleotide diversity and D estimates together, most PAV genes appear to be evolving under relaxed constraints.

A signature of relaxed selection associated with PAV genes is combined with a variety of features which have been associated with lower gene essentiality. We found that PAV genes have lower expression levels and higher tissue specificity; both of these features have been associated with higher rates of substitutions and reduced gene essentiality [112, 190]. Older genes have been considered more essential [18] and have been associated (in humans, flies and *Aspergillus*) with a higher expression level and stronger purifying selection [19]. We found that PAV genes are, on average, newer additions to the genome and that most exons affected by PAV do not have an orthologous exon in *A. lyrata* (663/794). We note that both E-PAV and CDS-PAV genes are enriched in reverse transcriptase domains (Figures 3.5 and 3.6) and E-PAV genes for transposase domains (Figure 3.5), suggesting exonization of transposable elements as the origin of some PAV-affected exons.

In addition, the fact that gene expression is only marginally reduced in accessions affected by exon deletion events suggests that the lost exons may only have had a limited impact on gene functionality. This is possibly explained in some cases by alternative splicing, which has already been associated with an increased frequency of exon loss in humans, mice and rats – alternatively spliced forms are less likely to be conserved between species than constitutive exons [201]. In *A. thaliana*, we found that genes with E-PAV are under weaker purifying selection and have a greater number of alternative splice events compared to intact genes. This observation suggests that alternatively spliced exons are likely to be under reduced selective constraints compared to constitutive exons and thus whole exon deletions would have less of a detrimental effect than the loss of a constitutive exon. To the best of our knowledge this is the first time that exon loss events have been associated with elevated alternative splicing levels within a species rather than between species.

The genomic context of genes has also been linked to both patterns of sequence evolution and features associated with gene essentiality. A recent study in *A. thaliana* has correlated the presence of TEs adjacent to genes with sequence variation within that gene [202] suggesting TEs tend to accumulate near genes under lower selective pressures located in regions with less efficient purging of TE sequence. Indeed, for our set of E-PAV genes, we find a higher density of TEs in the vicinity. In addition, we also find that genes undergoing PAV have an increased proportion of motifs associated with recombination hotspots within their sequence. Both findings are consistent with PAV events being associated with genes located in genomic regions evolving under reduced selective constraints. Moreover, higher TE content and hotspot motifs are consistent with the suggestion that unequal recombination between homologues may be a major mechanism for generating P/A polymorphisms [43]. However, it should be noted that no recombinogenic motif is both necessary and sufficient for a recombination event to occur [203] and as such, their

connection, if any, to PAV remains speculative. It is also worth noting that a higher TE density in particular regions can result in genome assembly artefacts [204] and that as such some neighbouring PAV events may be spurious.

Nevertheless, all of these features considered together suggest that although some individual deletions might have an adaptive value, overall coding region loss disproportionally affects genes under reduced selective pressures. So how are these results reconciled with the enrichment of certain gene families and GO functional terms? The enrichment of specific functional categories and gene families among PAV genes (Figure 3.2), leads to the implication of adaptive pressures favouring PAV on genes related to specific biological processes [43]. However, as we have shown, PAV genes are associated with a variety of features suggestive of lower selective constraints. We argue that the enrichment of certain GO categories and/or gene families among genes associated with a particular genomic feature does not, by itself, allow us to draw conclusions about any adaptive processes these genes may be undergoing. Consistent with this, we find that intact genes associated with the gene categories in which PAV genes are enriched, also show the same signatures of reduced selection (Table S3.6). This is notable for those sets of genes involved in, e.g., signal transduction, nucleic acid binding and the NBS-LRR family – categories enriched among PAV genes (Figure 3.2). For instance, if we compare the set of E-PAV genes to the set of genes with all exons present, and the set of NBS-LRR genes to the set of genes belonging to other families, we find that both E-PAV and NBS-LRR genes are comparatively newer additions to the genome, have a higher dN/dS ratio, a higher number of alternative splicing events, a higher number of paralogues, a higher proportion of SNPs and are found closer to TEs (Table S3.6). Note that genes with a higher dN/dS will have fewer instances of detectable homology across longer phylogenetic distances and as such their appearance of being 'relatively new' may be artefactual [20]. We also note that the proportion of polymorphic sites is higher not only in PAV genes but in genes of that functional category. To demonstrate that PAV genes do not bias the comparison of, e.g., the set of NBS-LRR genes to the set of genes belonging to other families, we repeat the analysis restricted to intact genes only and observe the same result (Table S3.6).

The fact that we observed fewer PAV genes than a previous study examining 80 fully sequenced *Arabidopsis* genomes ($n = 2741$ [43]) is likely due to differences in methodology. Firstly, our analysis uses 17 genomes assembled using a combination of read-to-reference genome (Col-0) alignment and *de novo* approaches, and – importantly – for which transcriptome data was available [39], rather than the 80 accessions reported by [40]. Secondly, we use a more conservative methodology for defining significant deletions whilst [43] define PAV genes using what is referred to as the 'broad definition': "one being found at a particular locus only in some genomes compared to the others." This allows a gene to be called as a PAV gene even if a copy exists at a different locus. To minimise the inclusion of rearrangement events as deletions, [43] examined their predicted PAV genes using blastn against a reference accession, excluding from the 'absent' category any gene with a counterpart that matches >50% of its length. Our definition

of PAV is more restrictive as we only deemed an exon or gene to be deleted if genome alignments showed that the deletion spanned at least a whole exon or whole gene with not a single identifiable base remaining. Finally, the [43] study used genomes assembled according to the TAIR8 annotated positions whereas our data is assembled according to TAIR10. There is a small risk, therefore, of having incorporated now-obsolete gene models into their findings. Regardless of the methodological differences and the resulting variation in sample size it is worth noting our results are not in contradiction to those of previous studies examining PAV both in *Arabidopsis* and other species as we find similar deviations from random expectations in the functional annotations of genes. Our analysis of sequence evolution and other genic features of PAV genes do not rule out the possibilities of conditional neutrality at PAV loci or that balancing selection may be acting on PAV genes, allowing adaptation to the environmental conditions of specific ecotypes. Instead, the findings presented show that PAV events can be explained by a non-adaptive interpretation where genes under reduced constraints are more susceptible to the spread of allele variants containing significant deletions.

In summary, our results suggest that although significant enrichment in functional categories among PAV genes was observed, most exon loss events are observed in newer, poorly functionally characterised genes associated with signatures linked to less essential genes evolving under lower purifying or balancing selection. This may reduce the potential functional relevance of structural variations within these genes. We conclude that while an adaptive model for PAV cannot be ruled out, the observed functional enrichments among PAV genes and increased nucleotide diversity can also be interpreted without invoking selection.

3.5 Materials and Methods

3.5.1 Genome sequence and annotations

Exon coordinates for *A. thaliana* strain Col-0 were obtained from The *Arabidopsis* Information Resource (ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR10_genome_release/TAIR10_gff3/TAIR10_GFF3_genes.gff, dated 20th March 2012). The genomes of 17 *A. thaliana* accessions (Bur-0, Can-0, Ct-1, Edi-0, Hi-0, Kn-0, Ler-0, Mt-0, No-0, Oy-0, Rsch-4, Sf-2, Tsu-0, Wil-2, Ws-0, Wu-0 and Zu-0) were obtained from [39]. We did not use data from Po-0 because it has an unusually high frequency of heterozygosity and high similarity to Oy-0 [39]. Each genome has been fully sequenced and assembled, using a combination of *de novo* assembly and read mapping to the reference accession, Col-0. Reads were mapped iteratively and at each iteration a consensus sequence derived. This process was repeated until additional rounds of iteration added fewer than 2% of the variation detected in the first iteration. At this point, any remaining variation in the sequence was considered ambiguous and thus excluded, attributable to residual heterozy-

gosity, repetitive read mappings which could not be resolved, or to copy number variation.

3.5.2 Detecting missing exons relative to Col-0

For this analysis we selected a set of deletions spanning at least one full exon in at least one accession relative to the Col-0 reference genome from a wider set of deletion events described by Gan *et al.* [39]. Exons absent in the Col-0 reference genome but present in other accessions are not included in any analysis. Confirmation of these deletions is described by the original authors who analysed deletion breakpoints [39]. In this dataset, deletion breakpoints were estimated to within ~30bp, with left and right consensus sequences established by growing inwards from these estimates using the read mapping information. If there was a deletion, these two ends would overlap. Gan *et al.* [39] confirmed this with alignments of the left and right consensus sequences, thus excluding errors of sequencing or misassembly. We further confirmed the presence or absence of each individual exon in each of 17 accessions relative to the Col-0 genome annotation using blastn with default parameters [158]. Sequence alignments were obtained using the best hit homologue and the Smith-Waterman algorithm (fasta35 with parameters –a –A) [159]. We confirmed an exon as missing if both (a) an alignment could not be made, and (b) if none of the nucleotide positions in the Col-0 exon mapped to any nucleotide in the accession.

3.5.3 Functional category enrichment analysis

Four gene classification schemes were obtained. 'GOslim' terms were obtained from The *Arabidopsis* Information Resource (ftp://ftp.arabidopsis.org/home/tair/Ontologies/Gene_Ontology/ATH_GO_GOSLIM.txt, dated 9th July 2013), excluding terms unsupported by experimental or computational analysis, i.e. evidence codes ND (no biological data available), NR (not recorded) and NAS (non-traceable author statement). 'GO' term annotations were obtained from Ensembl BioMart (17th July 2013) [205]. 'Pfam' terms were obtained from Pfam v27.0 (17th July 2013) [206]. In addition, 7119 genes were classified into 49 distinct families as in [39]. Statistical significance of the enrichment of both GOslim, GO terms, of Pfam class and family membership among both E-PAV and CDS-PAV-affected genes was assessed using a Monte Carlo random sampling (1000 randomisations), with the p-value of the enrichment of each category obtained using a Z-test. The significance of individual categories was corrected for multiple testing by the Benjamini-Hochberg procedure.

3.5.4 Sequence evolution analysis

To approximate selective constraint on a gene, we calculated both dN/dS and a neutraliy index, NI (see below). For each gene, we obtained a local alignment of the Col-0

primary transcript CDS against its *A. lyrata* orthologue, using the Smith-Waterman algorithm (fasta35 with parameters –a –A) [159]. dN/dS was calculated using the Yang and Nielson model, as implemented in the yn00 package of PAML [170]. Using substitution estimates, as above, and SNP data from [39], we also estimated Tajima's D [200] per gene. Nucleotide diversity is calculated according to [39]. The neutrality index for each sequence, NI, was calculated as $log((2D_s+1)(2P_n+1)/(2D_n+1)(2P_s+1))$, where D_n and D_s are the numbers of non-silent and silent substitutions, and P_n and P_s are the numbers of non-silent and silent polymorphisms [162]. NI can be interpreted in the same manner as a McDonald-Kreitman test for comparing the ratio of fixed to within-species differences: its symmetrical distribution allows the inference of purifying selection when NI > 0 and the inference of positive selection when NI < 0 [163].

3.5.5 Gene expression

Expression specificity was calculated as a tissue specificity index (*tau*) [149], using the parallel signature sequencing (MPSS) database [167, 168, 169]. Expression levels were calculated using RNA-seq transcript abundance data, as absolute read values corrected by gene length in each accession (known as rpkm values: per gene, the number of reads per kilobase per million mapped reads) [39].

3.5.6 Paralogue number and gene age annotations

Orthologue and paralogue data were obtained from BioMart [207]. A proxy for gene age was established using taxonomic classifications, based on the phylostratigraphic method of [208]. If a candidate orthologue was identified for each *A. thaliana* gene in any of 15 plant and algal species at a minimum identity of 30%, the gene was considered to be as old as the 'broadest' taxonomic category held in common (Table 3.2). This allowed us to make use of orthologue data despite divergence times relative to *A. thaliana* being known for only its closest relatives – at approximately 5 million years for *A. lyrata* [209], and 20 million years for *Brassica rapa* [170].

Age category	Species	Taxonomic classification (Kingdom, Division, Clade, Group, Order, Family, Genus, Species)	
		Shared	Differing at...
1	*Arabidopsis lyrata* (Northern rock cress)	Genus (*Arabidopsis*)	Species (*lyrata* vs. *thaliana*)
2	*Brassica rapa* (turnip mustard)	Family (Brassicaceae)	Genus (*Brassica* vs. *Arabidopsis*)
3	*Glycine max* (soybean)	Group (rosids)	Order (Fabales vs. Brassicales)
4	*Brachypodium distachyon* (purple false brome)	Division (Angiospermae)	Clade (monocot vs. eudicot)
4	*Oryza glaberrima* (African rice)	Division (Angiospermae)	Clade (monocot vs. eudicot)
4	*Oryza sativa* (Asian rice)	Division (Angiospermae)	Clade (monocot vs. eudicot)
4	*Populus trichocarpa* (black cottonwood)	Division (Angiospermae)	Clade (monocot vs. eudicot)
4	*Sorghum bicolor* (sorghum)	Division (Angiospermae)	Clade (monocot vs. eudicot)
4	*Zea mays* (maize)	Division (Angiospermae)	Clade (monocot vs. eudicot)
5	*Physcomitrella patens* (moss)	Kingdom (Plantae)	Division (Bryophyta vs. Angiospermae)
5	*Selaginella moellendorffii* (spike moss)	Kingdom (Plantae)	Division (Lycopodiophyta vs. Angiospermae)
5	*Vitis vinifera* (grape vine)	Kingdom (Plantae)	Division (Magnoliophyta vs. Angiospermae)
6	*Chlamydomonas reihhardtii* (green algae)	-	Kingdom (Protista vs. Plantae)
6	*Cyanidioschyzon merolae* (red algae)	-	Kingdom (Protista vs. Plantae)

Table 3.2: Age categories for orthologues of *A. thaliana* genes.

3.5.7 Transposable element and hotspot motif density

Transposable element (TE) coordinates for *A. thaliana* strain Col-0 were obtained from The *Arabidopsis* Information Resource (file 'TAIR10_Transposable_Elements.txt', dated 20th March 2012). For our analyses, we identified every instance of all 25 hotspot-associated motifs (of 5 to 9bp) described by [192] in the Col-0 reference genome. TE and hotspot motif density for each gene was calculated as the proportion of base pairs occupied by a TE or a hotspot motif within windows of size 1-100kb centred on the nucleotide at the gene's midpoint. Windows consist of both coding and non-coding sequence within a region of length (window size)/2 up- and downstream of the midpoint base. Both TE and hotspot motif density were calculated as the number of TE or motif bases, respectively, relative to the number of intergenic or genic bases contained within the window, rather than the total number of bases in the window.

3.5.8 Alternative splicing events

Alternative splicing events were identified using the methods described in [165]. In brief, the number of alternative splicing events per gene were identified by aligning EST data obtained from dbEST [164] to the genome sequence (ftp://ftp.ncbi.nih.gov/repository/dbEST, downloaded 1st May 2011). Those ESTs aligning to regions with no annotated gene were excluded from the analysis. EST alignments were then used to create an exon template. Alternative splicing events per gene were identified by comparing alignment coordinates for each individual EST to exon annotations. As a low EST coverage can increase the number of falsely positive claims that an exon is constitutive, rather than spliced, we excluded genes with 10 or fewer ESTs. ESTs were assigned to genes using gene annotation coordinates. A comparable alternative splicing index that avoids transcript coverage biases was obtained using the transcript normalisation method described

in [98]. Briefly, for each gene one hundred random samples of 10 ESTs were selected. Finally, the number of alternative splicing events were calculated for each random sample (as detailed above), with an overall average calculated per gene.

3.5.9 Randomisation test

A randomisation test was used to obtain numerical p values to assess the statistical significance of any variation in the characteristics of PAV-affected genes compared to 'intact' genes. In brief, we contrasted genomic feature parameters in E-PAV (n=330) or CDS-PAV genes ($n = 81$) to the distribution of means of the same genomic feature in $s = 10,000$ randomly generated subsets of an equal number of genes drawn from the complete gene set. The numerical p value was calculated as follows: let q be the number of times the mean value of the PAV set exceeded the mean value of the randomly generated subset. Letting $r = s - q$, then the p-value of this test is $(r+1)/(s+1)$.

Chapter 4

Correcting for differential transcript coverage reveals a strong relationship between alternative splicing and organism complexity

4.1 Summary

What at the genomic level underlies organism complexity? Although several genomic features have been associated with organism complexity, in the case of alternative splicing, which has long been proposed to explain the variation in complexity, no such link has been established. Here we analysed over 39 million ESTs available for 47 eukaryotic species with fully sequenced genomes to obtain a comparable index of alternative splicing estimates which corrects for the distorting effect of a variable number of transcripts per species – an important obstacle for comparative studies of alternative splicing. We find that alternative splicing has steadily increased over the last 1400 million years of eukaryotic evolution and is strongly associated with organism complexity, assayed as the number of cell types. Importantly, this association is not explained as a by-product of covariance between alternative splicing with other variables previously linked to complexity including gene content, protein length, proteome disorder and protein interactivity. In addition, we found no evidence to suggest that the relationship of alternative splicing to cell type number is explained by drift due to reduced N_e in more complex species. Taken together, our results firmly establish alternative splicing as a significant predictor of organism complexity and are, in principle, consistent with an important role of transcript diversification through alternative splicing as a means of determining a genome's functional information capacity.

4.2 Introduction

Prior to widespread genome sequencing, it was assumed that organism complexity was proportional to gene content – that more complex organisms encode a greater amount of genetic information [210], the unit of which is the gene [211]. However, the sequencing of the human genome, revealing a lower than expected number of genes [212], initiated a hunt to uncover the genomic basis of organism complexity [71] as, despite two rounds of whole genome duplication at the base of the vertebrate lineage [213, 214], the human genome contains almost as many genes as that of a worm [215]. Several genomic features have been shown to have a significant association with organism complexity, measured as the number of distinct cell types per species (cell type number; CTN). These variables include various measures of the potential number of molecular interactions per protein: the number and proportion of protein-protein interaction domains in each protein [99, 216] and protein disorder (flexibility in a protein's 3D structure to adopt a variety of conformations) [99, 217, 218]. More recently, total coding region length in a genome was shown to be positively associated with organism complexity [99]. This same study also showed that when restricting the analysis to metazoans, gene number becomes a significant predictor of organism complexity.

Alternative splicing, a post-transcriptional process in eukaryotes by which multiple distinct transcripts are produced from a single gene, has the potential to boost the total number of distinct proteins encoded in a genome in the absence of increases in gene number [71]. As such, an association between alternative splicing and organism complexity has long been proposed. Under an 'adaptive' model, an increase in alternative splicing could facilitate the evolution of higher organismal complexity, by increasing proteome diversity (and thus, diversifying functionality) at a level disproportionate to increases in the number of protein-coding genes [69, 70, 219]. Indeed, over the last decade, alternative splicing prevalence (the proportion of multi-exon genes that have at least one alternative splicing event) has been successively revised upwards for humans, with recent deep sequencing transcriptome analyses estimating that up to 94% of multi-exon genes undergo alternative splicing [220, 221]. However, assessing the expansion of alternative splicing prevalence through evolutionary time and establishing a link between alternative splicing and organism complexity have proved difficult [71]. The main barrier to comparative studies of alternative splicing prevalence arises from the fact that differences in transcript sequence coverage across species can distort both the proportion of genes classified as undergoing alternative splicing and the number of alternative splicing events detected [98, 222, 71, 99, 223, 224, 225]. Kim *et al.* [98] devised a method of transcript number normalisation to obtain comparable alternative splicing prevalence indices involving the identification of alternative splicing events from a random sample of 10 transcripts per gene. Importantly, they showed that alternative splicing in vertebrate species was higher than among invertebrates and that this was not explained by the higher abundance of transcripts available for vertebrate species. Although not directly tested, these findings were

suggestive of a link between alternative splicing and complexity as vertebrates are generally considered to have a higher CTN compared to invertebrates. Surprisingly, there are still no current datasets for comparable alternative splicing indices and controlling for transcript abundance in comparative analyses of alternative splicing prevalence is the exception rather than the rule. The resulting lack of comparable estimates for the number of alternative splicing events per gene has hampered efforts to quantify alternative splicing prevalence across taxa [226], the accumulation of splicing events over time [227] and the link between alternative splicing rates and organism complexity [71, 228]. The only attempt to directly assess the relationship between alternative splicing variation and CTN [99] was considered inconclusive by the authors because of the lack of comparable alternative splicing measures.

Here we assess the prevalence of alternative splicing in 47 eukaryotic genomes by calculating a comparable index of alternative splicing which corrects for differences in transcript coverage (adapted from [98]; see Materials and Methods). The species examined include metazoans, plants, fungi and protists. We then examined how these alternative splicing indices relate to organism complexity and compared the strength of alternative splicing as a predictor of CTN to previously described correlates, including the number of protein-interacting domains encoded per gene [216], protein disorder [217, 218, 99, 228], the number of protein-protein interactions, gene number and various measures of coding region length [99].

We find that alternative splicing has steadily increased over the last 1400 million years of eukaryotic evolution. We also find that alternative splicing is strongly associated with CTN and that this relationship is not a by-product of the relationship between various genomic features and complexity.

It is important to note that if increases in the proportion of alternatively spliced genes or the level of alternative splicing these genes undergo are linked with CTN, such an association would not constitute proof of causality. Under a 'non-adaptive' model, the association of alternative splicing and organism complexity could be a by-product of the link between complexity and a lower effective population size (N_e). The passive emergence of 'genomic complexity' and even organismal complexity itself is suggested by the work of Lynch and colleagues, who argue that non-adaptive processes explain the majority of the variance in organism complexity as 'more complex' organisms have a smaller N_e [73, 72]. As documented consequences of a comparatively small N_e include the accumulation of slightly deleterious mutations, both in coding [229, 47, 48] and regulatory [46] sequences, as well as an increase in average intron and coding region lengths [73], it is reasonable to expect that mutations impairing splicing regulation will accumulate more rapidly in 'more complex' organisms resulting in higher (but not necessarily functional) transcript diversity. Consistent with this, single species studies have shown that a significant proportion of alternative splicing events are probably the result of non-coding 'noise' and not biologically meaningful [74, 76].

Using a limited sample size, we do not find any evidence to suggest that the association of alternative splicing and CTN is explained by differences in N_e. To the best of our knowledge this is the most comprehensive assessment of alternative splicing levels covering all major eukaryotic taxa, and the first time in which the link between alternative splicing and CTN has been assessed using a comparative index of alternative splicing which corrects for differential transcript coverage.

4.3 Results

4.3.1 Alternative splicing prevalence has increased throughout evolutionary time

In order to assess whether alternative splicing levels have changed over time, over 39 million publicly available partial transcripts, representing 112 eukaryotes (20 protists, 18 plants, 23 fungi and 51 metazoans including 23 chordates), were aligned to their corresponding genomes to identify alternative splicing events (see Materials and Methods). To minimise the strong dependence of alternative splicing event detection on transcript coverage per gene [98, 222, 224, 71, 99, 223, 225], we used a transcript normalisation protocol [98] where alternative splicing events are identified in randomly selected samples of 10 ESTs per gene. We obtained a comparable alternative splicing index per gene by averaging the number of alternative splicing events in 100 samples [98] (Figure 4.1).

Using the comparable alternative splicing index, we calculated for each species both alternative splicing prevalence (ASP), defined as the proportion of alternatively spliced genes in the sample of genes analysed, and alternative splicing level (ASL), defined as the average number of alternative splicing events per gene. Genomes with comparable alternative splicing estimates available for fewer than 500 genes were excluded from further analyses leaving, in total, 47 species (6 protists, 10 plants, 6 fungi and 25 metazoans; Table S4.1). We found that both ASP and ASL vary among eukaryotic clades with chordates having both the highest ASP and ASL compared to non-chordate metazoans, fungi, plants and protists (Figure 4.2 and Table S4.1). While our ASP estimates are higher in most clades compared to a previous study based on eight species using comparable alternative splicing indices, the relative differences among clades are consistent [98].

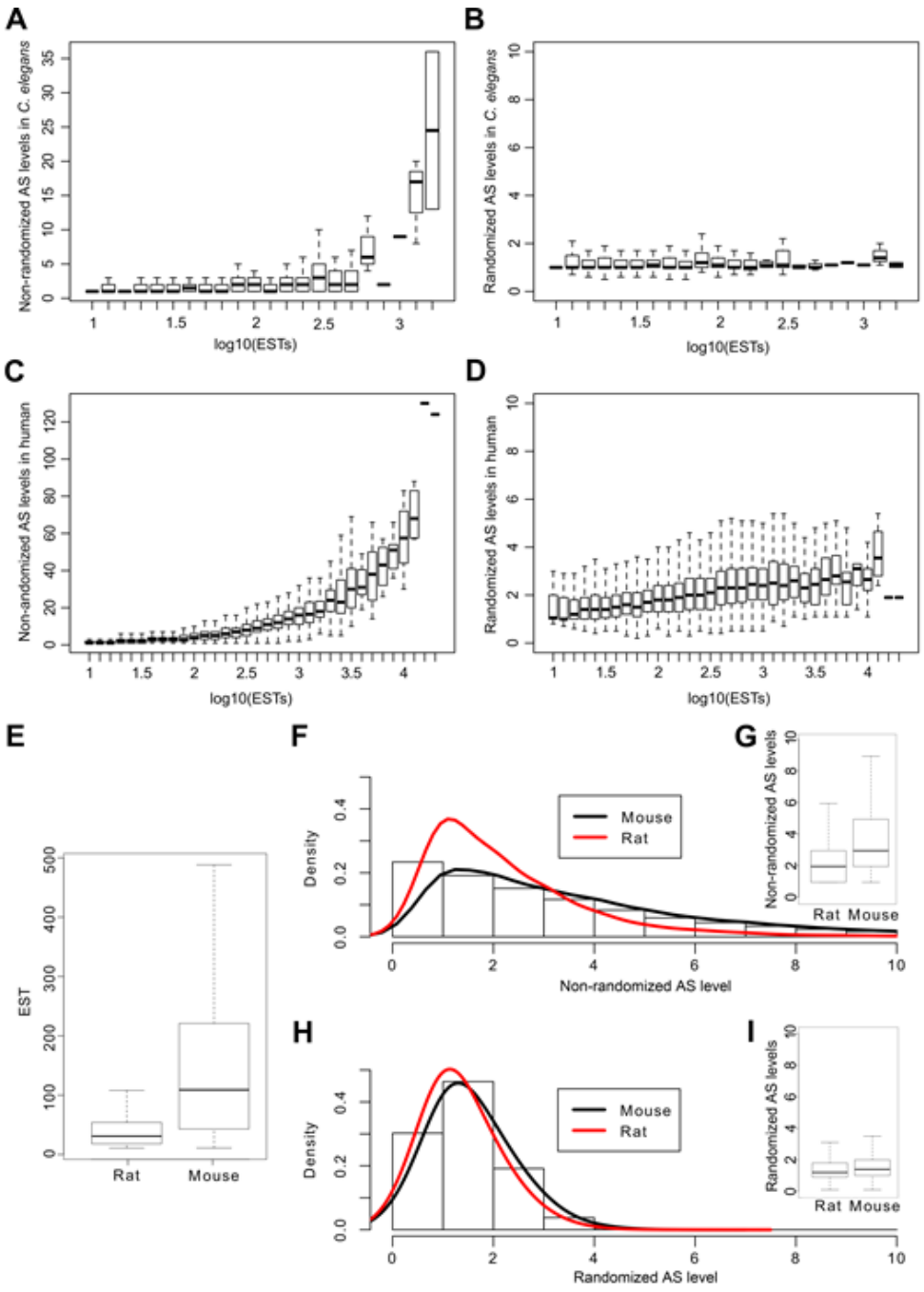

Figure 4.1: Total transcript number influences alternative splicing level (ASL) detection but this bias can be corrected using a sampling method.
ASL detection in genes divided by transcript coverage is shown for the nematode (A and B) and human (C and D) using both the full transcript dataset (A and C) and the random sampling method (B and D). Large differences in the average EST coverage for both rat and mouse (E) lead to correspondingly large differences in ASE detection for the two species (F). These are greatly reduced by the use of a sampling method (H). Inset panels G and I show the average ASL in both species using both the full transcript dataset and the random sampling method, respectively.

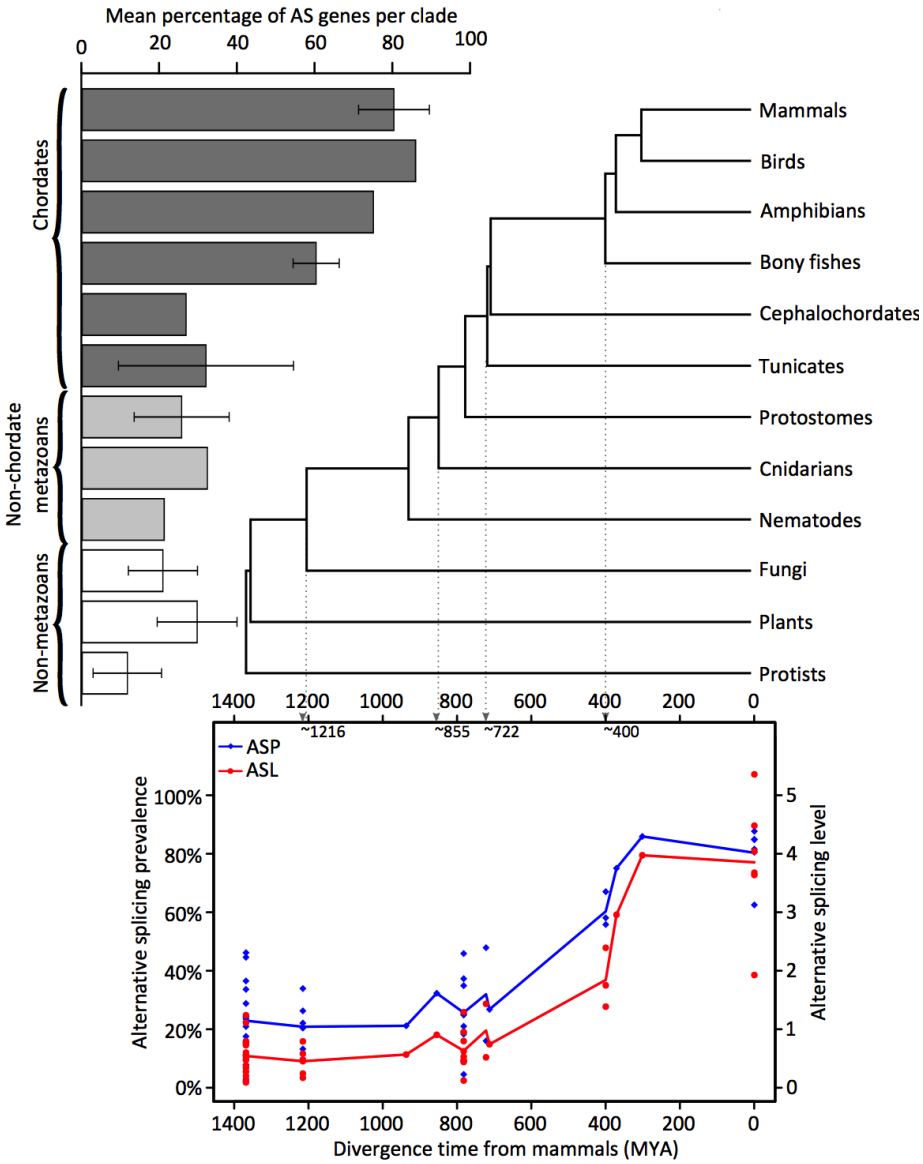

Figure 4.2: Variance in alternative splicing over evolutionary time.
Bars show the average percentage of alternatively spliced genes per species grouped according to their divergence from humans, as shown in the adjacent phylogenetic tree (data from [230]), and their taxonomic category (chordate, non-chordate metazoan, or non-metazoan). Note that categories without an error bar have only one member. The scatter plot shows changes in alternative splicing prevalance, i.e. the percentage of alternatively spliced genes per genome (blue) and in alternative splicing level, i.e. the average number of alternative splicing events per gene for each species (red). Trend lines represent the mean of all values at each divergence time. Although the relative positions of cephalochordates and tunicates on this tree are disputed [231], this does not significantly alter the trend.

An increase in alternative splicing through evolutionary time (Figure 4.2) is consistent with observations reporting links between ASP and evolutionary time restricted to metazoan species [227] and show that it is not an artifact of differential transcript coverage among species [71, 99]. The higher prevalence and levels of alternative splicing in plant species compared to fungi and protists suggest that AS levels have independently increased in this lineage.

Overall, by using comparable alternative splicing estimates from species covering all major eukaryotic clades and correcting for differential transcript coverage, we show that alternative splicing has increased over the last 1400 million years of eukaryotic evolution in the metazoan lineage with a more moderate and potentially independent rise in alternative splicing in plants.

4.3.2 Alternative splicing is a strong predictor of organism complexity, assayed as cell type diversity

A previous attempt to assess the link between alternative splicing and organism complexity, assayed as the number of distinct cell types [99], was rendered inconclusive because of the known bias caused by differential transcript sequence coverage among genes and species [98, 222, 224, 71, 99, 223, 225]. As such, we assessed the relationship of ASP and ASL with the number of distinct cell types per species (cell type number; CTN) as a proxy of organism complexity using the comparable AS index (see Materials and Methods). We found that both ASL and ASP are strongly associated with CTN (ASP: $r^2 = 0.76$, $p = 9.36 \times 10^{-9}$; ASL: $r^2 = 0.83$, $p = 1.77 \times 10^{-10}$; Table S4.2 and Figure 4.3). This association remains strong when restricting the analyses to the metazoan-fungi lineage (for ASP, $r^2 = 0.71$, $p = 2.45 \times 10^{-5}$, and for ASL, $r^2 = 0.81$, $p = 1.28 \times 10^{-6}$; Table S4.3).

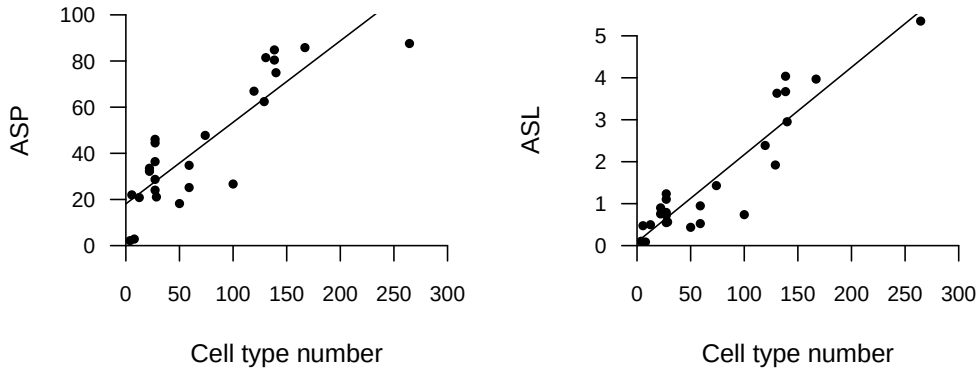

Figure 4.3: Relationship between alternative splicing and organism complexity, assayed as cell type number.
Graphs show the relationship between cell type number and ASP ($r^2 = 0.76$; $p = 9.36 \times 10^{-9}$) and ASL ($r^2 = 0.83$; $p = 1.77 \times 10^{-10}$).

Several genomic and functional parameters have previously been associated with organism complexity (using CTN as a proxy). Xia *et al.* [216] reported a strong link between CTN and protein-protein interaction (PPI) domain coverage. Other genomic variables found to have a more moderate association with CTN include protein disorder [217, 218, 99, 228] and proteome size (assayed as concatenated protein length) [99].

Gene number, previously found to be unrelated to CTN, has recently been reconsidered as a significant predictor but only after plant genomes are excluded from the analyses [99].

How does alternative splicing compare to these previously reported predictors of CTN? To address this, we compared the relationship between CTN and alternative splicing with that of 12 additional genomic measures of protein interactivity as well as proteome disorder, gene length and number, all previously linked to CTN (see Materials and Methods for descriptions and sources of each variable assessed). Of all parameters tested, ASL was found to have the strongest association with CTN ($r^2 = 0.83$, $p = 1.77 \times 10^{-10}$) followed by ASP and the average number of PPI domains per protein ($r^2 = 0.76$, $p = 9.36 \times 10^{-9}$ and $r^2 = 0.64$, $p = 8.19 \times 10^{-11}$ respectively; Table S4.2). We then re-examined the relationship between each parameter with CTN restricting the analyses to a set of 24 species for which data in all variables tested was available. The mean number of interactions per protein was not included in this or subsequent analyses due to the small number of species for which data was available ($n = 10$). ASL remained the top predictor of CTN ($r^2 = 0.87$, $p = 2.80 \times 10^{-11}$) with ASP showing an increased ($r^2 = 0.80$, $p = 2.66 \times 10^{-9}$) and the average number of PPI domains per protein a decreased association with CTN ($r^2 = 0.59$, $p = 6.42 \times 10^{-6}$; Table 4.1).

Category	Variable	Linear regression		Phylogenetic generalized least squares regression		
		r^2	p	r^2	p	λ
ALTERNATIVE SPLICING	Alternative splicing level	0.87	2.80×10^{-11}	0.87	1.59×10^{-13}	0
	Alternative splicing prevalence	0.80	2.66×10^{-9}	0.77	8.38×10^{-11}	0.05
SIZES AND LENGTHS	Number of genes	-0.01	0.40	0.26	1.23×10^{-3}	0.76
	Avg. protein length	-0.05	0.97	0.12	0.03	0.79
	Proteome information content	3.25×10^{-3}	0.31	0.09	0.05	0.65
	Proteome size	0.31	2.59×10^{-3}	0.49	4.08×10^{-6}	0.75
DISORDER	Mean % of disordered binding sites	-0.03	0.59	0.02	0.26	0.71
	Mean number of disordered binding sites	-0.04	0.78	-0.04	0.99	0.68
	Total number of disordered binding sites	0.04	0.18	0.21	3.97×10^{-3}	0.69
	Mean proteome disorder	-0.03	0.64	6.45×10^{-3}	0.34	0.71
INTERACTIVITY	% PPI domain seq per protein	0.60	5.36×10^{-6}	0.60	1.30×10^{-7}	0
	Avg. num of PPI domains per protein	0.59	6.42×10^{-6}	0.59	1.61×10^{-7}	0
	Proportion of proteins with 1+ PPI domains	0.54	2.33×10^{-5}	0.54	7.80×10^{-7}	0

Table 4.1: Association between CTN and genomic features before and after phylogenetic signal correction in 24 eukaryotic species.

As the relationship between genomic parameters and CTN has been shown to increase after the removal of plant genomes [99], we reassessed the predictive power of all parameters after restricting the analyses to the metazoan-fungi lineage. This resulted in a stronger association between CTN and many parameters with the two alternative splicing indices remaining the best predictors of CTN (Table S4.3). Consistent with previous findings [99], when plant genomes are excluded, gene number was found to be significantly associated with CTN ($r^2 = 0.34$, $p = 1.74 \times 10^{-3}$; Table S4.3).

Due to the tendency of related species to resemble one another it is also necessary to control for this non-independence in a comparative analysis of patterns across species.

Pagel's λ measures the extent to which observed correlations between traits reflect their shared evolutionary history assuming an evolutionary model under Brownian motion [232]. For the 24 species for which data in all variables tested was available, we obtained estimates of λ and restricted log-likelihood for the correlations between CTN and each genomic variables, re-calculating each correlation to account for phylogenetic non-independence of the variables by fitting a phylogenetic generalized least squares (PGLS) model (see Materials and Methods). ASL remained the top predictor of CTN even after taking into account the strength of the phylogenetic signal ($r^2 = 0.87$, $p = 1.59 \times 10^{-13}$, $\lambda = 0$), followed by ASP ($r^2 = 0.77$, $p = 8.38 \times 10^{-11}$, $\lambda = 0.052$) and the percentage of PPI domain sequence per protein ($r^2 = 0.60$, $p = 1.30 \times 10^{-7}$, $\lambda = 0$; Table 4.1). This pattern holds true if we only take into account metazoan and fungal species (Table S4.3).

As most of the assessed parameters co-vary among themselves (Tables S4.4 and S4.5), the association between some variables with CTN may be secondary to their covariance with another genomic feature which is in turn linked to CTN. In order to identify the genomic parameters which significantly contribute to CTN, we carried out a stepwise analysis (see Materials and Methods). In the optimal stepwise regression model, the majority of the variance in CTN is explained by ASL, supported by proteome size (Table 4.2). Similar results are obtained when constraining the data to the metazoan-fungal lineage (Table 4.2). In fact, contrasting each variable directly against AS by including ASL/ASP in multiple regression models with each additional variable revealed that in all cases only the AS parameter remained significantly associated with CTN (Table S4.2). The only exception was proteome size which remained significantly associated with CTN after correcting for either ASP or ASL but only when fungi and metazoans were included in the analysis (Table S4.3).

Data	Step	Variable added	r^2	Total r^2	p
All species (n=24)	1	ASL	0.87	0.87	2.72×10^{-11}
	2	Proteome size	0.02	0.88	0.05
Fungi-metazoan lineage (n=15)	1	ASL	0.86	0.86	4.78×10^{-7}
	2	Proteome size	0.04	0.89	1.80×10^{-5}

Table 4.2: Forward stepwise regression analysis using 13 functional genomic variables as predictors of CTN.
Note that due to limited data, the mean number of interactions per protein is not included as a variable in this analysis.

In order to best capture the predictive value of sets of co-varying variables, we used a principal component analysis to reduce the dimensionality among the 13 predictors of complexity. This analysis was performed on a subset of species where data was available for all predictors ($n = 24$). Interestingly, PC1 and PC2 (which explain 35.2% and 31.4% of the variance in the matrix, respectively) allow chordates to be differentiated from all

other species (Figure 4.4). Of all resulting principal components, we found that PC1 is the only significant predictor of CTN ($r^2 = 0.052$, $p = 8.58 \times 10^{-7}$). The two alternative splicing variables (ASP and ASL) and the three protein interactivity variables (average number of PPI domains per protein, PPI domain coverage and the proportion of proteins with at least one PPI domain) were found to be the main contributors to PC1. Similar results were obtained when restricting the analyses to the metazoan-fungi lineage (data not shown). It is worth noting, however, that the value of r^2 when regressing PC1 against CTN, when including either all species or only metazoans and fungi, is lower than that of ASL ($r^2 = 0.83$, $p = 1.77 \times 10^{-10}$), suggesting that collapsing the dimensionality of the variables does not improve the prediction of CTN beyond the variance explained by ASL alone.

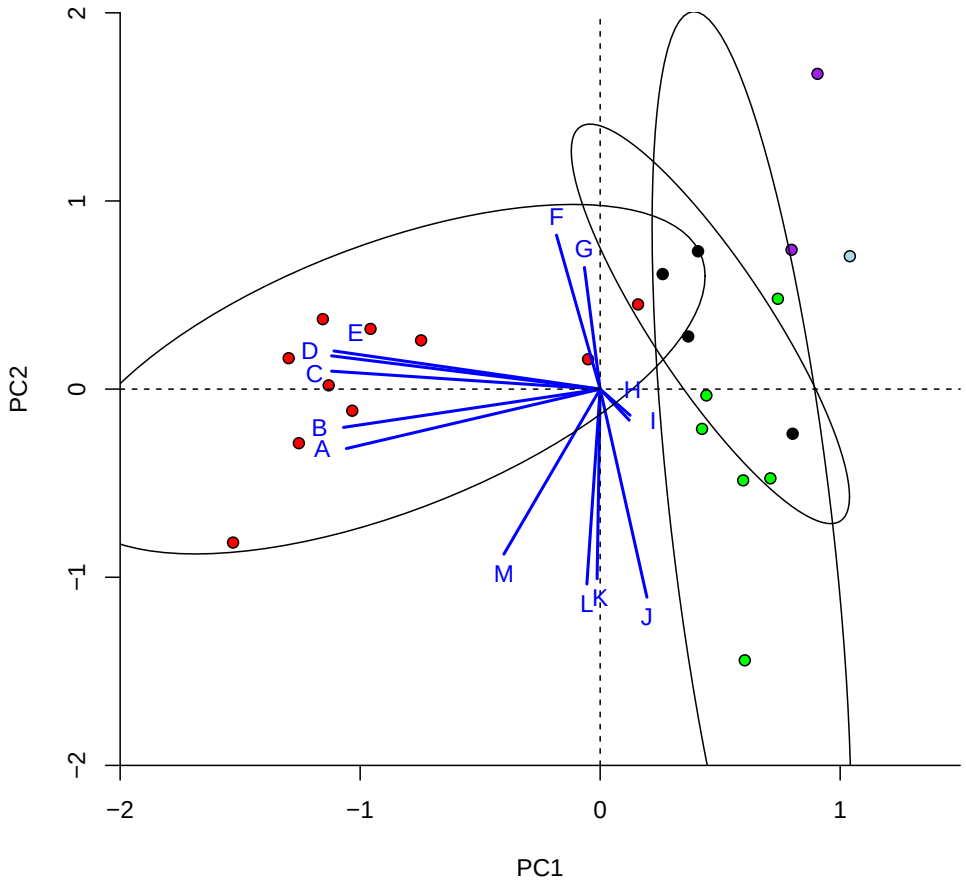

Figure 4.4: Biplot of the first two principal components built from 13 functional genomic variables available for 24 species.
Graph shows the distribution of species along PC1, which explains 35.2% of the variance in this dataset, and PC2, which accounts for 31.4%. Points represent each of 24 species for which data was available for all variables and are coloured by taxonomic category: chordates (red), non-chordate metazoans (black), plants (green), fungi (blue) and protists (purple). Ellipses show the clustering of species according to their dispersion along PC1 and PC2 (with confidence limit 0.95). Blue lines radiating from (0,0) represent each variable included in the analysis. The direction of each line represents the highest correlation coefficient between the PC scores and the variable, with the length of each line proportional to the strength of this correlation. Letter codes for each variable: ASL (A), ASP (B), % PPI domain sequence per protein (C), proportion of proteins with at least one PPI domain (D), avg. num. of PPI domains per protein (E), avg. protein length (F), mean number of disordered binding sites per protein (G), mean proteome disorder (H), mean % of disordered binding sites per protein (I), number of genes (J), total number of disordered binding sites per proteome (K), proteome information content (L) and proteome size (M). Refer to Table S4.1 for raw data.

The above results show that AS is significantly associated with CTN and that this association is not explained as a by-product of the relationship between AS and other genomic features also related to CTN. However, it is possible that some of these associations might be explained by ascertainment bias resulting from the fact that humans and other closely

related species have been disproportionately studied. With the exceptions of *C. elegans* and *D. melanogaster*, larger amounts of data exist for vertebrates than other species. It is possible that the higher estimates of AS and other genomic features, and even higher cell type number among vertebrates, might partly result from the greater availability of data for these species. To address this possibility, we used the total number of ESTs per species as a proxy for interest in a species as higher transcript availability has a direct impact on the quality of genome annotation. Compared to other proxies of 'research interest' such as 'number of publications per species', the number of ESTs approximates how much data has accumulated rather than how many interpretations of it there have been.

We established that the number of ESTs per species is significantly associated with various genomic characteristics (Table 4.3). Notably, ASL and ASP, as well as CTN, were found to be significantly related with transcript number per species (ASL: $r^2 = 0.45$, $p = 7.29 \times 10^{-7}$; ASP: $r^2 = 0.39$, $p = 8.01 \times 10^{-6}$; complexity $r^2 = 0.41$, $p = 5.01 \times 10^{-5}$).

Category	Variable	r^2	p	n
ALTERNATIVE SPLICING	Alternative splicing level	0.45	7.29×10^{-7}	47
	Alternative splicing prevalence	0.39	8.01×10^{-6}	47
SIZES	Avg. protein length	-3.97×10^{-3}	0.43	81
	Number of genes	0.09	2.48×10^{-3}	112
	Proteome information content	0.32	1.32×10^{-7}	80
	Proteome size	0.49	1.07×10^{-12}	81
DISORDER	Mean % of disordered binding sites	9.29×10^{-3}	0.33	35
	Mean number of disordered binding sites	-0.03	0.59	35
	Total number of disordered binding sites	0.24	4.31×10^{-3}	35
	Mean proteome disorder	-3.51×10^{-3}	0.40	35
INTERACTIVITY	% PPI domain seq per protein	0.14	1.21×10^{-3}	81
	Avg. num of PPI domains per protein	0.08	0.02	81
	Proportion of proteins with 1+ PPI domains	0.08	0.01	81
COMPLEXITY	Number of cell types	0.41	5.01×10^{-5}	37

Table 4.3: Regression coefficients of various functional parameters with number of ESTs as independent variable (quadratic polynomial regression).

Thus, we re-examined the relationship of CTN with AS and other gene features using the residuals of a quadratic polynomial regression with EST number. This correction resulted in a marked reduction in the variance in CTN explained by ASL and ASP ($r^2 = 0.47$, $p = 9.84 \times 10^{-5}$ and $r^2 = 0.57$, $p = 8.82 \times 10^{-6}$ respectively; Table S4.6). Correcting all variables by transcript coverage also reduced the predictive value of other gene features for CTN (Table S4.6). However, the relative order of gene feature parameters as predictors of CTN remained unaltered with splicing and, to a lesser extent, the degree of protein-protein interactivity the most strongly associated with CTN (Table S4.6). Furthermore, if considering all 13 variables, the optimal stepwise regression model (see Materials and Methods) explained 90% of the variance in CTN ($p = 1.81 \times 10^{-5}$), with the strongest of five predictors being ASP (Table 4.4). When restricting the analyses to the fungi-metazoan lineage we found that the optimal regression model contained only

two regressors, ASP and the mean percentage of disordered binding sites per protein (see Materials and Methods for a description of this variable) (Table 4.4). In fact, only three parameters (average protein length, the number of genes and the total number of disordered binding sites per protein) remained significantly associated with CTN in a regression model directly comparing each variable with either ASP or ASL (Table S4.5). An alternative transformation of the data, taking the natural log of EST number, resulted in lower correlation coefficients but the relative strength of each variable in a regression against complexity remained unchanged (Table S4.7).

Data	Step	Variable added	r^2	Total r^2	p
All species (n=24)	1	ASP	0.67	0.67	5.82×10^{-7}
	2	Number of genes	0.09	0.76	5.22×10^{-3}
	3	Avg. num. PPI domains per protein	0.08	0.84	2.61×10^{-3}
	4	Proteome information content	0.04	0.88	0.02
	5	Avg. protein length	0.02	0.90	0.03
Fungi-metazoan lineage (n=15)	1	ASP	0.88	0.88	1.55×10^{-7}
	2	Mean % of disordered binding sites	0.05	0.93	0.01

Table 4.4: Regression coefficients of 14 genomic parameters with CTN as the dependent variable, using as estimates for each variable the residuals of a linear regression between variable x and the log-transformed number of ESTs per species.
Note that due to limited data, the mean number of interactions per protein is not included as a variable in this analysis.

Our data spans a diverse range of species with associated variations in the number of available ESTs per species (Table S4.1). For genomes with lower EST numbers (often those that also have a lower CTN), highly expressed genes will make a disproportionate contribution to each species' comparative alternative splicing index as the number of genes with the minimum required number of ESTs will be smaller. As such, we expect lowly expressed genes to primarily contribute data for genomes with a higher number of available ESTs.

Under the 'non-adaptive' model, a reduced N_e among 'more complex' organisms (assayed as those with higher CTN) would result in an accumulation of mutations detrimental to splicing regulation, potentially resulting in the proliferation of 'noisy' alternative splicing events. Such neutral increases in alternative splicing should be particularly pronounced among lowly expressed genes which, on average, are under lower selective pressures compared to highly expressed genes. Importantly for this study, if lowly expressed genes are more highly spliced, then our data would overestimate ASL for species with high EST numbers, artificially inflating the correlation strength with CTN.

Using microarray data for four model species (human, mouse, *C. elegans* and *A. thaliana*; see Materials and Methods), we find that, as expected, there is a strong correlation between the number of ESTs per gene and gene expression level. However, contrary to the prediction of the non-adaptive model, we found that the more highly expressed genes have

a higher number of alternative splicing events (Figures 4.5, 4.6, 4.7 and 4.8). Therefore, our data might be underestimating ASP and ASL in genomes with a higher number of available ESTs, as more lowly expressed genes – with lower alternative splicing levels – disproportionately contribute to the species' alternative splicing indices. By extension, the relationship of AS with CTN might also be underestimated.

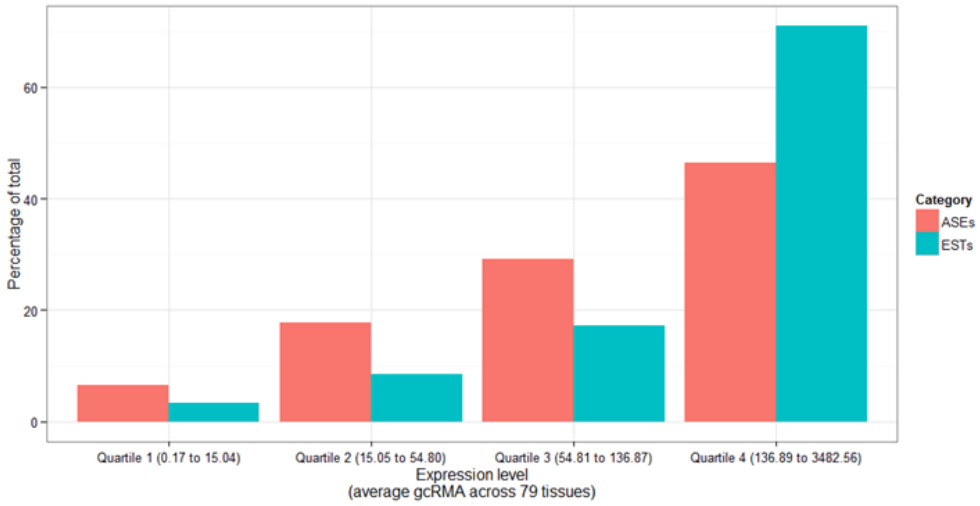

Figure 4.5: Distribution of both alternative splicing events ($n = 17,738$) and ESTs ($n = 646,634$) for the genome of *A. thaliana*, according to the expression level of the associated gene.
Expression data is partitioned into quartiles each of size $n = 4579$. Correlations of the number of alternative splicing events and the number of ESTs against expression level are significant in both cases (Spearman's $rho = 0.29$, $p < 2.2 \times 10^{-16}$ and $rho = 0.81$, $p < 2.2 \times 10^{-16}$).

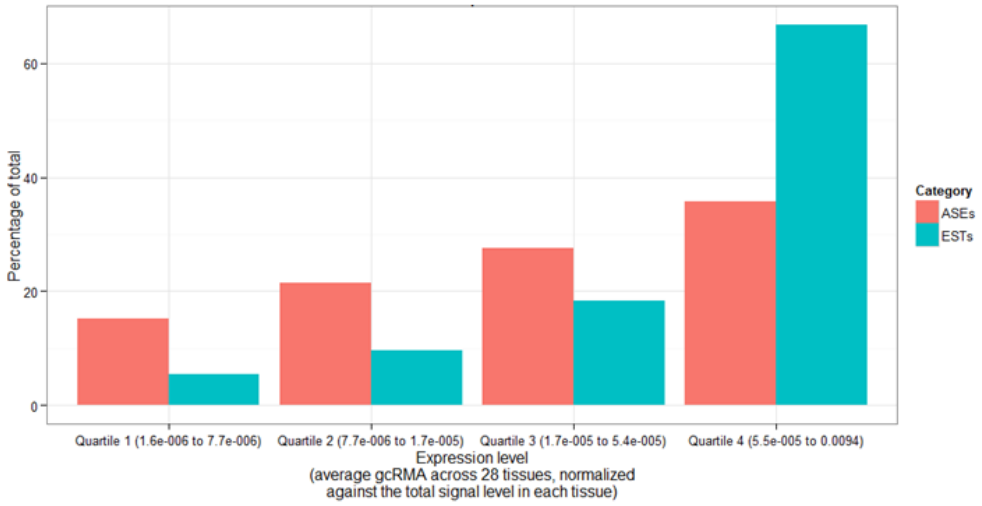

Figure 4.6: Distribution of both alternative splicing events ($n = 76,699$) and ESTs ($n = 4,510,520$) for the genome of *H. sapiens*, according to the expression level of the associated gene.
Expression data is partitioned into quartiles each of size $n = 1618$. Correlations of the number of alternative splicing events and the number of ESTs against expression level are significant in both cases (Spearman's $rho = 0.35$, $p < 2.2 \times 10^{-16}$ and $rho = 0.53$, $p < 2.2 \times 10^{-16}$).

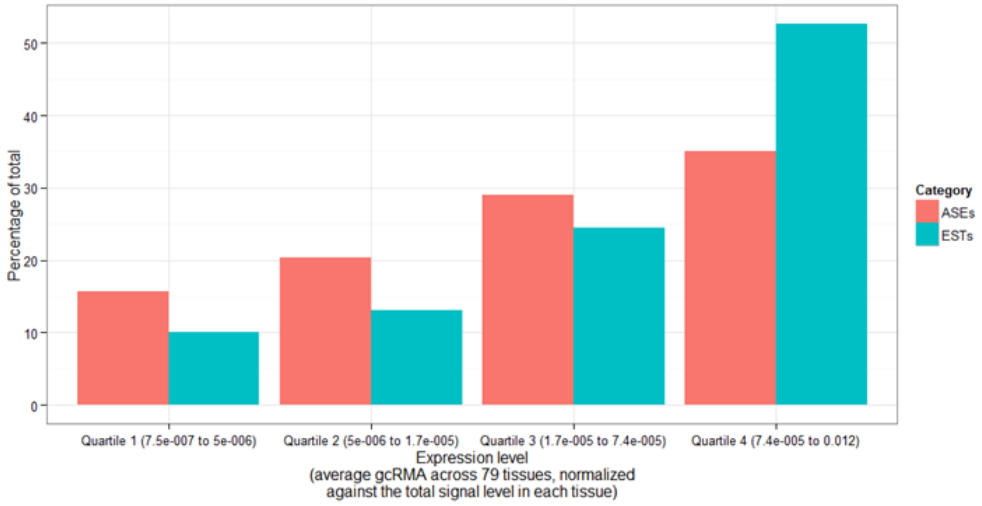

Figure 4.7: Distribution of both alternative splicing events ($n = 45,628$) and ESTs ($n = 1,403,152$) for the genome of *M. musculus*, according to the expression level of the associated gene.
Expression data is partitioned into quartiles each of size $n = 1812$. Correlations of the number of alternative splicing events and the number of ESTs against expression level are significant in both cases (Spearman's $rho = 0.32$, $p < 2.2 \times 10^{-16}$ and $rho = 0.51$, $p < 2.2 \times 10^{-16}$).

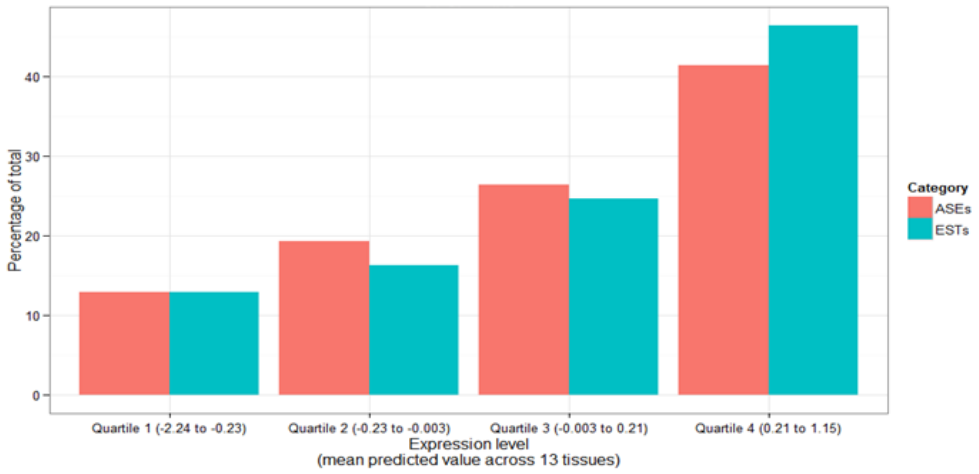

Figure 4.8: Distribution of both alternative splicing events ($n = 2128$) and ESTs ($n = 80,237$) for the genome of *C. elegans*, according to the expression level of the associated gene.
Expression data is partitioned into quartiles each of size $n = 2690$. Correlations of the number of alternative splicing events and the number of ESTs against expression level are significant in both cases (Spearman's $rho = 0.81$, $p < 2.2 \times 10^{-16}$ and $rho = 0.29$, $p < 2.2 \times 10^{-16}$).

4.4 Discussion

Here we have assessed alternative splicing levels in 47 eukaryotic species and showed that alternative splicing has increased over the last 1400 million years of evolution. Our data ranges from *P. falciparum*, in which 3% of genes are spliced with an average of 0.09 splice events per gene, to humans, where 88% of genes are spliced with an average of 5.35 splice events per gene. Consistent with the findings of Kim *et al.* [98], we find that chordates have higher levels of alternative splicing than any other taxonomic group with mammals and birds having both proportionately more genes that are alternatively spliced (ASP) and a higher number of alternative splicing events per gene (ASL). We observed significant increases over time in ASP and ASL for the opisthokonts, and show that past claims for an increased level of alternative splicing along the evolution of metazoans are not explained by differential transcript coverage [227]. Our data do not support a previous claim for lower alternative splicing levels among birds compared to mammalian species [233] and in fact, alternative splicing levels in the chicken genome were found to be among the highest of all species tested.

Plant genomes were found to have higher levels of alternative splicing than both protist and fungal species, comparable to those found among invertebrate species. This is consistent with relatively low levels of alternative splicing in the eukaryotic ancestor with independent rises in the plant and metazoan lineages. None of the plant genomes we examined, however, match the levels of alternative splicing observed in the vertebrate lineage.

Our results demonstrate a strong association between alternative splicing and organism complexity providing, to the best of our knowledge, the first systematic evidence for a link between these two variables. In this study we have used the number of cell types as a proxy for organism complexity. Cell type number has been proposed as an indicator of an organism complexity as the higher number of components or cell types in 'more complex' organisms should reflect, to some degree, their higher number of functions [234]. We acknowledge, however, that complexity is difficult to define and even more difficult to measure and that all operational definitions for 'complexity' are, to various degrees, contentious [235]. Several proxies of organismal complexity have been proposed, however these measures are either relevant to some taxonomic groups, such as encephalization coefficient, or no measurements are available for a large number of species, such as phenotypic complexity [236]. While accepting that 'organism complexity' is likely to be a multidimensional variable encompassing many other features, we chose this measure as, compared to other proxies, cell types are more easily quantifiable for organisms from distant taxonomic groups. It is important to note that, as CTN data are drawn from a diverse range of studies (see Materials and Methods), more detailed characterisations of CTN can appear anomalous. For example, we expect chimpanzees to have a similar CTN to humans but currently, humans are the better characterised species and as such the human CTN appears higher (Table S4.1). To address whether this type of outlier confounds our results, we repeat our analyses using the average CTN for the order each species belongs to. This makes the assumption that any variation in CTN between species of a given order reflects measurement noise, rather than relevant biological information. Our results do not significantly differ when using these alternate values of CTN (Tables S4.8 and S4.9).

Importantly, as most past studies analysing the relationship between various genomic features and organism complexity have adopted cell type number as a proxy [237, 99, 216, 228], its use allowed us to contrast our results with those of others. Such comparisons showed that the relationship of alternative splicing and CTN is not secondary to other genomic features previously associated with CTN, including proteome size (measured as total protein coding sequence length [99]), protein disorder [228, 99], and protein interactivity.

Before the full sequencing of nuclear eukaryotic genomes became widespread, gene number was expected to have a direct relationship with organism complexity as more genes would encode a higher number of proteins boosting the number of potential molecular interactions [218, 217]. The sequencing of the human genome, however, found no evidence for such an association [212]. The discrepancy between organism complexity and gene content became known as the G-paradox [238, 210, 239, 240]. However, a recent study concluded that gene number and organism complexity are related after all, albeit only when plant species are removed from the analyses [99].

Our findings also support a significant association between gene number and CTN in the absence of plant genomes ($r^2 = 0.34$, $p = 1.74 \times 10^{-3}$; Table S4.3). However, alternative

splicing level has a stronger association with CTN ($r^2 = 0.77$, $p = 1.09 \times 10^{-8}$) and is sufficient to explain the relationship between CTN and gene number.

Unlike alternative splicing and gene number which directly impact on the number of interacting proteins, additional gene features linked to CTN can boost the interactivity potential of individual proteins without expanding their number. One of the simplest measures of the functional potential of the proteome, total coding region length, has been found to be significantly associated with CTN [99]. Although we observed a similar association between proteome size and CTN, this relationship is entirely explained as a by-product of both variables' covariance with alternative splicing. Proteome size remains a marginal, albeit significant, predictor of CTN in a stepwise regression model restricted to the metazoan and fungi lineage where alternative splicing level was the strongest variable (Table 4.1). Moreover, proteome size was not a significant contributor to the only principal component found to be significantly associated with CTN.

Protein disorder – the lack of equilibrium in a protein's 3D structure under physiological conditions [218] – has been proposed as a candidate predictor of organism complexity as higher intrinsic disorder allows individual proteins to adopt a greater variety of conformations, increasing the average number of interacting partners per protein and potentially boosting functional diversification of the proteome [218, 217]. Nevertheless, subsequent findings show the association between disorder and CTN only explains any substantial amount of variance when bacterial species are included [99, 228]. Our analyses of protein disorder using both stepwise regressions and principal component analysis do not provide any evidence of hidden covariance between protein disorder and CTN. Moreover, despite the fact that past studies have found a higher than expected number of disordered motifs in alternatively spliced areas at the gene level [241, 218] we do not find a significant association between protein disorder and alternative splicing per species (Tables S4.5 and S4.6).

Finally, a third measure of potential molecular interactions per protein, the presence of protein-protein interaction domains, has been shown to be strongly associated with CTN [216]. We found three protein interactivity parameters – PPI domain coverage, the average number of PPI domains per protein and the proportion of proteins with at least one PPI domain – to be significantly associated with CTN regardless of the set of species examined (Tables S4.2 and S4.3). A head-to-head comparison between predictors of CTN showed that protein interactivity measures are better predictors of CTN than any other variable with the exception of alternative splicing. After controlling for alternative splicing, however, no protein interactivity parameter was found to be significantly associated with CTN (Tables S4.2 and S4.3). An additional measure of protein interactivity previously associated with CTN, the mean number of protein-protein interactions [99], was not included in most of our analyses as data was limited to only 10 species in our set. These comparisons show that although protein interactivity is significantly associated with CTN there is a great overlap between the variance in CTN explained by protein interactivity

and that explained by alternative splicing.

Several studies have proposed an association between alternative splicing and protein domain content, suggesting that alternative splicing could act as a buffer against reduced functionality because of 'domain overload' – too many protein domains or domains in the wrong combination [242, 79, 243]. A large scale analysis has shown that protein domains are non-randomly combined in functional proteins with fewer protein domain co-occurrences observed than expected, suggesting that certain protein domains 'avoid' each other [244], whilst other domains – including PPI domains – are 'promiscuous' and tend to coexist within individual transcripts [245]. Our analyses of covariance among functional gene variables showed that alternative splicing and PPI measures are positively correlated – genomes with higher levels of alternative splicing also have a higher PPI domain presence. We further examined the association between ASL and PPI domain coverage within species but found only a marginal association between the two variables constrained to a few species (Table S4.10). This finding suggests that although genomes with a high level of alternative splicing also tend to have a higher PPI domain coverage, there is no support for a role for alternative splicing acting as a buffer of PPI domain overload.

Overall, our results are consistent with a direct association between alternative splicing and CTN, one which is not explained by other genomic features previously associated with organism complexity. This finding is, in principle, consistent with previous suggestions that alternative splicing may underlie the rise in complexity during eukaryotic evolution thanks to its potential to expand transcript diversity and thereby increase the number of potential molecular interactions and functions (reviewed in [71, 219, 69]).

Nevertheless, it is important to note that the rise in CTN has been accompanied by a reduction in effective population size [73]. Classical nearly neutral theory proposes that as effective population sizes diminish so too does the efficiency of purifying selection, resulting in the accumulation of slightly deleterious mutations, both in coding [229, 48, 47] and regulatory [46] sequences. The increased role of drift relative to selection has also been invoked to explain the proliferation of a number of genomic features among increasingly complex species [72, 73]. Although more recent studies have disputed this conclusion [246, 247, 248], a significant proportion of alternative splicing events have nevertheless been suggested to result from 'noisy' alternative splicing [74, 76, 77]. Thus, it is possible that the observed increase in alternative splicing among more complex species might be the result of increased genetic drift as a result of reductions in effective population size, rather than being directly associated with organism complexity. Using estimates of $N_e\mu$ (a composite parameter of effective population size and mutation rate) for the 12 species represented in this study [73] we found that a genome's capacity for alternative splicing remains strongly correlated with CTN even after controlling for $N_e\mu$ (partial Spearman's correlation coefficients: ASL$= 0.71,\ p = 2.37 \times 10^{-3}$; ASP$= 0.70,\ p = 3.35 \times 10^{-3}$). Although based on a small sample of species, this finding suggests that the association

between CTN and alternative splicing is not a by-product of reduced effective population sizes among more complex species. Future studies should be able to assess the functional contribution of increases in alternative splicing in the eukaryotic lineages we report here.

In addition, it is worth noting that a significant correlation of any genomic feature with CTN does not necessarily demonstrate a causal role on the evolution of organism complexity, i.e. a higher CTN. It is beyond the scope of this study to address this directly. Nevertheless, network theory provides some clues which allows us to speculate as to the likelihood that increases in transcript diversification, facilitated by alternative splicing, have affected the evolution of organism complexity. Boolean networks have been proposed as models for genetic networks as the attractors, representing different stable patterns of gene expression, correspond to different cell types [249, 250]. In Boolean networks, increases in the number of nodes leads to a higher number of attractors within the network at a rate equal to or exceeding the square root of the number of nodes in the network [251]. If we imagine each distinct transcript as a node in the genetic network we can speculate that alternative splicing, by increasing the number of nodes (transcripts), would lead to an increased number of attractors (cell types). Indeed, a previous study that generated relational networks for seven species associated the number of functions in a proteome with the number of polyform transcriptional units in the genome, those that produce protein isoforms with different functional assignments (which are strongly associated with the levels of splicing). Various properties of these networks (such as the number of nodes) were found to be strongly associated with organism complexity, suggesting a link between splicing and both multifunctionality and multicellularity [252].

We conclude that alternative splicing increases over the last 1400 million years of eukaryotic evolution are strongly associated with CTN. Furthermore, this association is stronger and more robust than other parameters previously associated with CTN although we cannot rule out the contributions of other genomic features as many co-vary. Our findings are consistent with an 'adaptive' scenario whereby a genome's capacity for alternative splicing – with its resulting expansion of the transcript pool – could constitute a critical component of the underlying mechanisms explaining the diversification of cell types and the rise in organism complexity over time. Nevertheless, the data here presented do not allow us to reach a conclusion on the functional relevance of increases in alternative splicing or to establish causality regarding the association of alternative splicing and organism complexity; thus, it is possible that a 'non-adaptive model' may account for it.

To the best of our knowledge, our results represent the first systematic assessment of the relationship between alternative splicing, evolutionary time and CTN and provide evidence for a strong association of alternative splicing and CTN. Our results further constitute the most comprehensive head-to-head comparison, to date, of variables associated with CTN.

4.5 Materials and Methods

4.5.1 Organism complexity

The number of unique cell types was used as a proxy of organism complexity. Estimates of cell type number per species were compiled from [253, 99, 254, 255, 256, 257]; data in graph form from [253] as interpreted by both [258] and [259] was also included. Following the methodology of [259], where more than one estimate of cell type number was available for a species, the average of the minimum and maximum number was used. In addition, we included a revised cell type number estimate for humans [260]. Table S4.1 provides averaged complexity estimates for both pro- and eukaryotic species whereas Table S4.11 shows the sources.

4.5.2 Identification of alternative splicing events

Comparable alternative splicing events were obtained using the following approach. Over 39 million EST sequences, accounting for 112 species, were downloaded from dbEST [164] and matched to their corresponding genome using GMAP [261] (these species are identified in Table S4.1 by a positive value in the column titled 'total number of ESTs'). Genome sequences and annotations were obtained from sources contained in Table S4.1. Cancer-derived EST libraries from human and mouse were removed from all analyses presented. To ensure high quality alignments, we only retained those ESTs with 95% identity. ESTs were assigned to genes using gene annotation coordinates. EST alignments were then used to create an exon template. These templates were generally in agreement with existing exon annotations but also identify a small number of non-annotated exons and discard orphan exons likely to be nested genes. Alternative splicing events per gene were identified by comparing alignment coordinates for each individual EST to exon annotations. A comparable alternative splicing index that avoids transcript coverage biases was obtained using the transcript normalisation method described in [98]. Briefly, for each gene with greater than 10 ESTs one hundred random samples of 10 ESTs were selected. The number of alternative splicing events were calculated for each random sample (as detailed above), with an overall average calculated per gene. The ability of this method to correct for transcript coverage bias and calculate an accurate number of alternative splicing events is shown in Figure 4.1. To estimate alternative splicing prevalence, a gene was considered to be alternatively spliced if it had at least an average of one alternative splicing event identified in each of the 100 random samples.

4.5.3 Additional functional genomic parameters

Gene number per species was obtained from Ensembl BioMart version 0.8 (March 2013) [262]. Proteome size (total amino acids encoded by all peptides), proteome information

content (total amino acids encoded by primary transcripts only) and average protein length were calculated from mRNA transcripts obtained from Ensembl BioMart version 0.8 (March 2013) [262]. The exception is the lancelet, Branchiostoma floridae, where transcripts were obtained from [263]. Protein-protein interaction (PPI) domains per protein were identified using HMMER3 with default parameters [264] and the Pfam-A database [265], with results parsed to consider matches to the 642 confirmed PPI domains as described in [216]. Protein disorder data was obtained from [99]. 'Disordered sites' are those which are not at equilibrium in the protein's 3D structure under physiological conditions and can thus adopt a greater variety of conformations. We obtained the mean number of disordered binding sites per protein, the total number of disordered binding sites across all annotated proteins per species, and the mean percentage of disordered binding sites per protein (i.e. the mean number of disordered sites per protein as a percentage of the protein's length). The latter term is considered the disorder of the protein. Mean proteome disorder is taken as the mean disorder per protein. The average number of protein-protein interactions per protein for each species was also obtained from [99]. Data on the composite parameter $N_e\mu$ (effective population size and mutation rate) was obtained from [73].

4.5.4 Statistical analysis

All statistical tests were performed in R, version 2.15.2 [266]. For stepwise regression analysis, new regressors are included at each step according to the Akaike Information Criterion [267], estimated using the package 'MASS' [90]. Principal components analysis was performed using the R packages 'FactoMineR' [268] and 'Vegan'.

4.5.5 Correction for phylogenetic autocorrelation

To assess and control for the strength of the phylogenetic signal on the correlation between CTN and the different genomic variables in this study, we computed Pagel's λ [232] based on maximization of the restricted log-likelihood using the gls subroutine from the R-package nlme [269]. Optimum negative values of Pagel's λ are reported as $\lambda = 0$. We used the subroutine PGLS in the R-package Caper [270] to examine the 'true' associations between the different genomic variables and CTN after using the optimal λ values to control for the strength of the phylogenetic signal. This method implements generalized least squares models which account for phylogenetic non-independence by incorporating the covariance between taxa into comparisons that determine the correlation between dependent and independent variables. PGLS is an extension of the independent contrasts methods proposed by Felsenstein [271] that provides a more general and flexible approach for assessing correlations between traits while accounting for phylogenetic divergence. An ultrametric phylogenetic tree for the analysed species was created by obtaining the divergence time between each pair of species from [230].

4.5.6 Expression level

Microarray data for four species (*H. sapiens*, *M. musculus*, *A. thaliana* and *C. elegans*) was obtained from the following sources. For *H. sapiens* and *M. musculus*, Affymetrix array data analysed by [272] was obtained from BioGPS (http://biogps.org). For *H. sapiens*, we obtained the expression of 11,449 genes across 28 tissues. We summarized gcRMA (GC robust multi-array average) normalized probe intensity levels to Ensembl IDs corresponding to protein coding genes. All probes matching to more than one Ensembl gene ID were removed. Expression values were then normalized against the total signal level in each tissue. For *M. musculus*, we obtained 9825 genes with one-to-one orthologues in the human across 79 different tissues and cell types. Where more than one array exists for a given tissue, data were averaged. The per probe expression signal was summarized to Ensembl gene IDs using the average expression of all the probe sets matching a single Ensembl ID. All probes matching to more than one Ensembl gene ID were removed. Expression values were then normalized against the total signal level in each tissue. For *A. thaliana*, data was obtained from the *Arabidopsis* Development Atlas (ADA), as generated by the AtGenExpress Consortium [62] (NASCARRAYS reference numbers 149-154, together representing 79 tissues, were downloaded from NASC AffyWatch (http://affymetrix.arabidopsis.info/)). Expression level was then quantified as the average gcRMA across all 79 tissues (with each value itself the mean of triplets) [108]. For *C. elegans*, tissue-specific expression for 13 tissues (germline, hypodermis, intestine, muscle, neurons, pharynx, coelomocytes, distal tip, excretory cells, spermatheca, spermatheca uterine valve, uterus and vulva) were obtained from [273] (http://wormtissue.princeton.edu), who analyzed a compendium of 916 microarray experiments from 53 datasets. Expression values in this dataset are already normalized to have mean 0 and variance 1. Expression level is taken as the mean across all tissues.

Chapter 5

General discussion

This thesis has addressed various questions relating to how signs of adaptive evolution can be distinguished from neutral evolution. A general conclusion of the work presented in this thesis is the importance of carefully contrasting any observations suggestive of positive selection (e.g. a disproportionate enrichment amongst certain functional categories of genes with particular structural variants) with a null hypothesis of neutral evolution [122]. This allows the following questions to be asked: firstly, how are potential signs of selection interpreted if these observations are considered in isolation (i.e. for individual genes or a set of genes of interest), and secondly, what can reasonably be inferred if instead there is an established context for any observations, e.g. data pertaining to the remaining genes in a genome? Taken together, the emphasis is upon what would falsify an adaptive hypothesis, a question more easily addressed given the ongoing expansion of available genomic data [104] and accordingly the means by which non-adaptive alternatives are evaluated. These points are addressed in all chapters and demonstrate that in some instances a neutral interpretation can explain observations that have been considered adaptive.

Chapter 2 discusses how individual signatures of selection may themselves be influenced by biasing factors. Various characteristics can be used to infer whether a gene is under selection. One of the most common is a sequence's dN/dS ratio, which quantifies the degree of substitution acting on both synonymous and non-synonymous sites, with high values (> 1) suggestive either of positive selection or relaxed selective constraint [102]. However, interpretation of this ratio in such a way as to infer the activity of selection requires certain assumptions. One of the key assumptions is that selective pressures upon synonymous sites are constant, such that dS (the number of synonymous substitutions per synonymous site) is an effective proxy of the background rate of mutation.

Chapter 2 demonstrates that when taking into account the higher level of purifying selection at exon edges, this doesn't, in general, mask signatures of positive selection, but does weaken the relationship of a gene's structural and functional characteristics to its dN/dS. This is of interest as if any structural or functional traits are the target of selection, the strength of selection acting explains less variance in the substitution rate than

expected. For instance, more highly expressed genes have their products more frequently 'exposed' to selection and accordingly the signature of selection (i.e. dN/dS) is expected to correlate with expression level. That exon edge conservation partially accounts for this relationship suggests that there is increased scope for the role of other evolutionary forces – including selection acting upon other parameters (such as splicing accuracy) – at explaining variation in substitution rate. This suggests future avenues for research, as it is of interest to understand which forces, acting upon which parameters, influence substitution rate. We have shown that when accounting the effect of conservation at exon edges, the ability of many genomic parameters at explaining variation in dN/dS is diminished (Table S2.12). We do not find that exon edge conservation masks a stronger association between dN/dS and any given parameter, as none explain greater variation in dN/dS after junction regions are removed. However, when accounting for lineage-specific substitution patterns, some genomic parameters do explain a greater amount of variation in dN/dS, in particular gene length (Table S2.17). This work is limited by the draft state of the *T. parvula* genome (used as an outgroup to calculate lineage-specific dN/dS for *A. thaliana*) and the accordingly small amount of data available ($n = 73$) to address the two factors – lineage-specific substitution and junction conservation – together. As such, the aggregate role of both factors in mitigating dN/dS biases is not yet known.

It would also be of interest to examine the relationship of gene length to lineage-specific dN/dS in further detail, after junction regions (with their higher selective constraint) are removed. We found that gene length is one of the strongest covariates of lineage-specific dN/dS, equivalent in strength to expression level, and that the relationship between the two is not explained by the covariance of length with expression level. If lineage-specific substitution rates indeed provide more accurate estimates of dN/dS, then junction removal – which removes a further source of bias from dN/dS – would better reveal the relationship of dN/dS to gene length.

If, as chapter 2 suggests, a gene's length and expression level are indeed among the stronger predictors of dN/dS, and the relationship of either parameter to dN/dS is not entirely explained by its covariance with the other, it is of interest to know what selection is acting upon, and why. For instance, although selection is likely to influence a gene's length and expression profile, it is unclear in what manner selection is acting – is a gene's length in part explained by it being selected for higher expression, or can genes become more highly expressed because their length is constrained?

It has already been shown that selection for translational efficiency can influence gene length – in animals, longer genes will be expressed at lower levels, assuming that elongation rate is the limiting step in translation [274]. Alternatively, if transcription and translation are slow, expensive processes in general, shorter genes will be biosynthetically economical and enriched amongst the highly expressed [126, 275]. Explanations underlying the relationship of gene length to expression are more speculative in plants, but are likely to involve a gene's intron content – both the total and average intron lengths

are longer in plants than in animals, with their regulatory roles perhaps necessary for higher levels of expression [276].

Overall, removing sources of bias from dN/dS estimates and reassessing the relationship of dN/dS to its known correlates – in particular, gene length and expression – will help identify which characteristics are under potential selection.

Chapter 3 examines a polymorphic variation within a single species. It had previously been shown that in *A. thaliana* genes with presence/absence variation (PAV) in their coding regions – i.e. structural variation that is potentially adaptive – are enriched for *R* (resistance) genes [184, 185], including members of the NBS-LRR family [43], known to be involved in pathogen detection [187]. Due to their recent, and substantial, structural divergence, alongside frequent recombination, copy number variation and high dN/dS ratios, *R* genes have been considered strong candidates for undergoing positive selection [41, 42]. It is reasonable to consider whether PAV, which contributes to this divergence, is adaptive also. However, functional categorisation of these PAV genes was defined without reference to the set of 'intact' genes in *A. thaliana*, and as such it is arguable that certain data – of use in clarifying the evolutionary history of PAV – is absent. This data can be considered 'contextual information', a broad term representing any comparative observations drawn from the wider genome, not just the set of genes of interest. This would allow the observed characteristics of any genes of interest – by definition, a subset of the wider genome – to be more accurately contrasted with the expected characteristics of the 'average gene.'

To that end, chapter 3 shows that the characteristics of genes with PAV in *A. thaliana* are similar to the characteristics of intact genes of that functional group anyway (Table S3.6). For example, the NBS-LRR family is enriched in genes with polymorphic exon deletions (exon presence/absence variation; E-PAV). Given the known functionality of, and the recent divergence amongst members of, the NBS-LRR family, E-PAV is arguably adaptive in this context. Nevertheless, we find the set of E-PAV genes, if compared with the set of genes with all exons present, and the set of NBS-LRR genes, if compared to the set of genes belonging to other families, have similar properties. The observation that NBS-LRR genes are enriched for E-PAV and the implication that this E-PAV is adaptive can then be re-interpreted: NBS-LRR genes, with or without E-PAV, are similar to genes in which E-PAV is observed anyway. The set of NBS-LRR genes and the set of E-PAV genes can both be characterised as being newer additions to the genome (i.e. are likely to be under relaxed selective constraint as they are less likely to be essential), having higher dN/dS ratios (i.e. consistent with a scenario of relaxed purifying selection), having a higher number of paralogues (i.e. there is a possible degree of functional redundancy in their protein products, such that mutations in one gene are less likely to be deleterious), and having a higher number of alternative splicing events (i.e. functional protein products encoded by the gene do not necessitate the presence of every exon). In this respect, we show that an observation interpretable as adaptive (E-PAV) can also be considered as

comparable to background selection.

Defining the characteristics common to a set of genes with shared function (NBS-LRR) and a set of genes with shared variation (PAV) suggests a direction for future research. As an *a posteriori* explanation of positive selection can reasonably be considered meaningful if significant variance is observed in particular functional categories (such as NBS-LRR), it is of interest to further characterise members of those categories in which the proportion of positively selected genes is enriched. It is reasonable to believe that genes in certain functional categories are more likely to experience positive selection due to the exposure of their protein products to strong and ongoing environmental pressures related directly to their function (e.g. those involved in sensory perception or immune responses [277], or those involved in the virulence of pathogens [278]). If a whole category is likely to have experienced greater positive selection there may be structural commonalities between those individual genes considered as positively selected and other members of that category. These other members may not show signs of positive selection at the point of observation but are more likely to have been influenced by it at some point in their evolutionary history. It would be of interest to identify how the structural characteristics of genes vary by how positively selected the 'average gene' in that functional group is likely to be. This could provide an evidence base for comparative purposes when candidate signatures of positive selection are found. Studies of positive selection often report their findings by functional category (e.g. positively selected genes in seven ant genomes are enriched for those with mitochondrial functions [279]; in the sea urchin *Allocentrotus fragilis*, for sulphur metabolism [280]), and in this respect categories enriched for positive selection are identified alongside individual genes.

Taken together, chapters 2 and 3 provide examples of how adaptive signatures can be distorted or re-interpreted, and that the potential signs of selection can be more reliably assessed against the null hypothesis of molecular evolution, neutrality [122], when an additional amount of 'contextual information' is available. Such 'context' can be the in-depth characterisation of a limited number of genomes or the broad characterisation of a single feature in many genomes.

In this respect, chapter 4 discusses the relationship of alternative splicing to organism complexity, assayed as the number of unique cell types, using data spanning all major eukaryotic taxa. Although a relationship has previously been proposed between these two variables, arguing that increased alternative splicing leads to greater proteome diversity – and thus, organism functionality – irrespective of increases in gene content [70, 219, 69], this has historically proved difficult to test [71]. One of the principal problems is that comparative estimates of the number of alternative splicing events per gene, and the proportion of genes that are alternatively spliced, are susceptible to biases introduced by the differential coverage of transcripts per species. This chapter describes and implements a comparable index of alternative splicing estimates to account for these transcript coverage biases, showing that the rise in alternative splicing over eukaryotic history is strongly

associated with increases in organism complexity. Furthermore, this association is not explained by covariance with other variables previously linked to complexity, such as proteome disorder [217, 218] and the degree of protein-protein interactions [216].

This is suggestive, although not explicit evidence of, an adaptive role of alternative splicing throughout eukaryotic evolution, and is consistent with previous reports ascribing a central role to alternative splicing in many biological processes (see chapter 1). It is also consistent with a previous report that genes with higher levels of alternative splicing are enriched for cytoskeleton-associated functions in 'complex' vertebrates [93], suggesting a role for alternative splicing in the evolution of cellular complexity. Although these observations are consistent with the expansion of functionally relevant alternative splicing throughout eukaryotic evolution, it is still contentious as to whether alternative splicing has been shaped largely by selective or by neutral forces.

It has been argued that as 'more complex' organisms have lower effective population sizes, the relative strength of drift to selection increases with organism complexity [72, 73]. As a consequence of this, there is a greater likelihood of slightly deleterious mutations – including those affecting splicing regulation – reaching fixation. This would result in a higher diversity of transcripts, but not necessarily those that are biologically meaningful. This is supported by the fact that few alternative splicing events appear conserved between 'higher' species (i.e. those with a higher number of cell types) and that as a consequence most alternative splicing is arguably 'transcriptome noise' [74, 76] – observations of such splicing events are consistent with their passive emergence via drift and their maintenance is best explained by the fact that purifying selection has yet to eliminate them. Even so, these studies contrast the presence of alternative splicing only amongst the 'most complex' species – humans and mice – and as such the absolute number of functionally relevant alternatively spliced genes, and the distribution of these genes by functional category, could still have been shaped by adaptive processes.

It would be of interest to identify the degree of conservation of alternative splicing events across the spectrum of species with comparatively lower levels of complexity (i.e. cell type number) to determine whether 'transcriptome noise' is indeed common throughout eukaryotic history, and whether – by extension – there are conserved functional categories in which alternative splicing is enriched. This could identify candidate groups within which any evidence of selection may be examined.

Although an association is found between alternative splicing and organism complexity, chapter 4 also shows that this is concomitant with – but not itself explained by – the decline in effective population size (N_e). Notably, however, this is not evidence of a causal relationship and as such the above 'non-adaptive' model of alternative splicing evolution may yet account for it. Nevertheless, this finding is limited by the small quantity of N_e data available ($n = 12$) [73] and a natural extension would be to employ a proxy, expanding the scope of the study and more reliably determining the contribution made by N_e. Although compendiums of N_e data have developed substantially [281, 282], there is

still minimal overlap between the set of species for which N_e is estimated and the set of species for which alternative splicing levels are estimated. As such, the most appropriate proxy would appear to be generation time, a known negative correlate of N_e [283]. Generation time data for 31 species for which alternative splicing estimates are also available can be found in previous studies [284, 285, 286, 287, 288, 289, 290, 291, 292, 293, 294, 295, 296, 297, 298, 299, 300]; compiling such a dataset would better address the extent to which neutral processes can explain the relationship of complexity and alternative splicing. Future studies can address other predictions of the 'non-adaptive' hypothesis for genomic complexity, particularly with regard to the evolution of exon/intron architecture. The scope of this hypothesis is broad, although largely theoretical, and as such a reasonable means of furthering chapter 4 would be to establish explicitly testable predictions with regard to quantifiable genomic characteristics. This could expand the evidence base for determining whether certain aspects of genome evolution – in particular, but not limited to, the rise in alternative splicing – are indeed primarily associated with neutral processes. Of particular interest is whether increases in the relative strength of drift to selection are associated with changes to the exon/intron architecture that are themselves associated with alternative splicing [301, 302, 303].

For instance, as a reduction in N_e also decreases the strength of selection against mildly deleterious into-intron insertions, gene length and intron content are expanded in 'more complex' genomes [73]. Importantly, intron size in vertebrates is a strong determinant of whether splicing proceeds via the exon definition or intron definition pathway – i.e. whether the splicing machinery is placed across the exon or the intron – with the latter pathway more efficient at shorter intron lengths, but becoming less so beyond approx. 250nt, whereupon the exon definition pathway becomes more common [304]. Further to this, we can predict that the lengthening of introns increases the selective pressure upon the (comparatively) smaller exons for stronger splice site signals – consistent with observations in *C. elegans* [305], *D. melanogaster* [306], fungi [307], humans and mice [308] – which could lead to a higher number of included exons. Nevertheless, the effect of intron length upon the relative strength of intronic and exonic splice site recognition may also lead to increased alternative splicing, particularly exon skipping, if exonic splice sites are, for instance, insufficiently strong. This would result in certain exons, within an increasingly intron-rich sea, being 'too weak' to distinguish from their flanking introns. This is consistent with observations in humans, mice and rats: compared to constitutive exons, alternatively spliced exons have higher dN/dS ratios, suggesting relaxed selective constraint with regard to their functional contribution to any protein product [309] and by extension, predicting they have weaker splice sites.

As such, there are both theoretical and empirical links between gene structure and alternative splicing, which emphasise the non-adaptive aspects of this association.

General trends in the evolution of exon/intron architecture – increases in intron relative to exon content (e.g. as reported by [310, 311, 153]) – are also consistent with the predictions

of this 'non-adaptive' model, and as such it is reasonable to expect that the observed increase in alternative splicing over eukaryotic history [152] naturally follows.

Observations consistent with these predictions would support the notion that the observed instances of alternative splicing are, in general, largely non-functional. By contrast, deviation from this neutral expectation could shed additional light on the coding potential of alternative splicing and any role it may have in proteome expansion, particularly if the extent to which genes undergo alternative splicing is disproportionately enriched for certain families. Additional predictions consistent with this non-adaptive hypothesis of splicing evolution can be made. We initially expect that as organisms with a lower N_e have a longer generation time [283], and, by association, a slower rate of molecular evolution (due to less frequent genome replication) [312], mutations affecting splicing regulation – which could lead to new splice variants – should be less likely to arise (as there are a lower number of DNA replication errors per unit time) although will be less likely to be eliminated if they do (as the strength of drift relative to selection is higher). We also expect the likelihood of such mutations to be increased by two other factors. Firstly, organisms with longer generation times are expected to 'tolerate' longer introns, as fewer genes will be under strong selection for mRNA processing time [288], and secondly, longer introns are more prone to splicing error, having a greater number of mutable sites than shorter introns [313]. As such, this predicts a higher proportion of genes to be alternatively spliced in organisms with a longer generation time.

In addition, as a gene's structure is inherently conducive to alternative splicing – those with a higher number of exons, for instance, can generate more alternatively spliced transcripts by probable combinations alone [314] – the number of alternative splicing events (ASEs) per gene should correlate with the number of possible ASEs per gene, with these correlation strengths themselves strongly associated with both N_e and generation time. For each gene with n exons, a maximum of $2n - 1$ transcript isoforms are possible assuming alternative splicing is limited only to distinct exon-intron combinations (i.e. excluding, for instance, alternative promoter or polyadenylation usage). There is no reason to believe that anything other than a limited proportion of this total would be observed, although it is reasonable to expect a strong correlation between the number of observed ASEs (i.e. ASL) and the number of possible ASEs. The hypothesis that eukaryotic splicing is predominantly noisy predicts that the strength of these correlations will be higher for species with lower N_e.

Taking these factors together, a reasonable null hypothesis is that the majority of ASEs are 'transcriptome noise' [76] – that they are primarily the consequence of a non-adaptive expansion of the eukaryotic exon/intron architecture and accordingly show few signs either of conservation or of functionality. Nevertheless, given the importance of alternative splicing to many biological processes (discussed in chapter 1), there arguably remains a role for selection in the evolutionary history of splicing, alongside the role of drift. As such, future work to characterise the exon/intron architectures of alternatively spliced

genes in multiple taxa, and to examine the evidence of conservation for alternatively spliced regions, may yet uncover adaptive signatures. For this example and in general, compiling and contrasting evidence of selection alongside evidence of non-adaptive processes is necessary to establish the evolutionary history of biological phenomena.

In summary, this thesis has addressed the means by which the activity of selection at the genome-wide level can be distinguished from neutral processes, focusing on three aspects of genome evolution: the rate of protein evolution, the distribution of presence/absence variation and the evolution of alternative splicing. A generalised conclusion is that the signatures of adaptation can either be masked by, or re-interpreted in the context of, non-adaptive processes and that with the increasing availability of high-throughput data, such considerations are of increasing relevance.

Bibliography

[1] T. Mitchell-Olds, J.H. Willis, and D.B. Goldstein. Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nature Reviews G*, 8(11):845–856, 2007.

[2] L.D. Hurst. Fundamental concepts in genetics: genetics and the understanding of selection. *Nature Reviews Genetics*, 10(2):83–93, 2009.

[3] Z. Yang and J.P. Bielawski. Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution*, 15(12):496–503, 2000.

[4] Lynda F. Delph and John K. Kelly. On the importance of balancing selection in plants. *New Phytologist*, 201(1):45–56, 2014.

[5] M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626, 1968.

[6] E. V. Koonin. A non-adaptationist perspective on evolution of genomic complexity or the continued dethroning of man. *Cell Cycle*, 3(3):280–285, 2004.

[7] A. Wagner. Neutralism and selectionism: a network-based reconciliation. *Nature Reviews. Genetics*, 9(12):965–974, 2008.

[8] M. Kimura. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*, 267(5608):275–276, 1977.

[9] Sankar Subramanian and David M. Lambert. Selective constraints determine the time dependency of molecular rates for human nuclear genomes. *Genome Biology and Evolution*, 4(11):1127–1132, 2012.

[10] S. Kryazhimskiy and J.B. Plotkin. The population genetics of dn/ds. *PLoS Genetics*, 4(12):e1000304, 2008.

[11] R. Nielsen and Z. Yang. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral dna. *Molecular Biology and Evolution*, 20(8):1231–1239, 2003.

[12] Jacob A. Tennessen. Positive selection drives a correlation between non-synonymous/synonymous divergence and functional divergence. *Bioinformatics*, 24(12):1421–1425, 2008.

[13] Eric J. Vallender and Bruce T. Lahn. Positive selection on the human genome. *Human Molecular Genetics*, 13(suppl 2):R245–R254, 2004.

[14] Pavlos Pavlidis, Jeffrey D. Jensen, Wolfgang Stephan, and Alexandros Stamatakis. A critical assessment of storytelling: Gene ontology categories and the importance of validating genomic scans. *Molecular Biology and Evolution*, 29(10):3237–3248, 2012.

[15] A. Decottignies, I. Sanchez-Perez, and P. Nurse. *Schizosaccharomyces pombe* essential genes: a pilot study. *Genome Res*, 13(3):399–406, 2003.

[16] A. G. Fraser, R. S. Kamath, P. Zipperlen, M. Martinez-Campos, M. Sohrmann, and J. Ahringer. Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference. *Nature*, 408(6810):325–30, 2000.

[17] C. Pal, B. Papp, and L. D. Hurst. Highly expressed genes in yeast evolve slowly. *Genetics*, 158(2):927–931, 2001.

[18] Wei-Hua Chen, Kalliopi Trachana, Martin J. Lercher, and Peer Bork. Younger genes are less likely to be essential than older genes, and duplicates are less likely to be essential than singletons of the same age. *Molecular Biology and Evolution*, 29(7):1703–1706, 2012.

[19] Yuri I. Wolf, Pavel S. Novichkov, Georgy P. Karev, Eugene V. Koonin, and David J. Lipman. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proceedings of the National Academy of Sciences of the United States of America*, 106(18):7273–7280, 2009.

[20] E. Elhaik, N. Sabath, and D. Graur. The "inverse relationship between evolutionary rate and age of mammalian genes" is an artifact of increased genetic distance with rate of evolution and time of divergence. *Molecular Biology and Evolution*, 23(1):1–3, 2006.

[21] Leonardo Arbiza, Joaquin Dopazo, and Hernan Dopazo. Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. *PLoS Computational Biology*, 2(4):e38, 2006.

[22] Margaret A. Bakewell, Peng Shi, and Jianzhi Zhang. More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proceedings of the National Academy of Sciences*, 104(18):7489–7494, 2007.

[23] Aristeidis Parmakelis, Marina Moustaka, Nikolaos Poulakakis, Christos Louis, Michel A. Slotman, Jonathon C. Marshall, Parfait H. Awono-Ambene, Christophe Antonio-Nkondjio, Frederic Simard, Adalgisa Caccone, and Jeffrey R. Powell. *Anopheles* immune genes and amino acid sites evolving under the effect of positive selection. *PLoS ONE*, 5(1):e8885, 2010.

[24] Macarena Toll-Riera, Steve Laurie, and M. Mar Alba. Lineage-specific variation in intensity of natural selection in mammals. *Molecular Biology and Evolution*, 28(1):383–398, 2011.

[25] Gareth D. Weedall, Spencer D. Polley, and David J. Conway. Gene-specific signatures of elevated non-synonymous substitution rates correlate poorly across the *Plasmodium* genus. *PLoS ONE*, 3(5):e2281, 2008.

[26] Susanta K. Behura and David W. Severson. Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes. *Biological Reviews*, 88(1):49–61, 2013.

[27] Maya Botzman and Hanah Margalit. Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. *Genome Biology*, 12(10):R109, 2011.

[28] L. Duret and N. Galtier. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual Review of Genomics and Human Genetics*, 10:285–311, 2009.

[29] N. Lartillot. Interaction between selection and biased gene conversion in mammalian protein-coding sequence evolution revealed by a phylogenetic covariance analysis. *Molecular Biology and Evolution*, 30(2):356–368, 2013.

[30] E. F. Caceres and L.D. Hurst. The evolution, impact and properties of exonic splice enhancers. *Genome*, 14:R143, 2013.

[31] Benjamin J Blencowe. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends in Biochemical Sciences*, 25(3):106 – 110, 2000.

[32] Roland Tacke and James L. Manley. Determinants of SR protein specificity. *Current Opinion in Cell Biology*, 11(3):358–362, 1999.

[33] Zhi-Ming Zheng. Regulation of alternative RNA splicing by exon definition and exon sequences in viral and mammalian gene expression. *Journal of Biomedical Science*, 11(3):278–294, 2004.

[34] A. L. Hughes. Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity (Edinb)*, 99(4):364–73, 2007.

[35] N. Stoletzki and A. Eyre-Walker. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. *Molecular Biology and Evolution*, 24(2):374–381, 2007.

[36] P.W. Messer and D.A. Petrov. Frequent adaptation and the McDonald-Kreitman test. *Proceedings of the National Academy of Sciences of the United States of America*, 110(21):8615–8620, 2013.

[37] Raquel Egea, Sonia Casillas, and Antonio Barbadilla. Standard and generalized McDonald-Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucleic Acids Research*, 36(suppl 2):W157–W162, 2008.

[38] Consortium Genomes Project, G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, and G. A. McVean. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73, 2010.

[39] Xiangchao Gan, Oliver Stegle, Jonas Behr, Joshua G. Steffen, Philipp Drewe, Katie L. Hildebrand, Rune Lyngsoe, Sebastian J. Schultheiss, Edward J. Osborne, Vipin T. Sreedharan, Andre Kahles, Regina Bohnert, Geraldine Jean, Paul Derwent, Paul Kersey, Eric J. Belfield, Nicholas P. Harberd, Eric Kemen, Christopher Toomajian, Paula X. Kover, Richard M. Clark, Gunnar Ratsch, and Richard Mott. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature*, 477(7365):419–423, 2011.

[40] J. Cao, K. Schneeberger, S. Ossowski, T. Gunther, S. Bender, J. Fitz, D. Koenig, C. Lanz, O. Stegle, C. Lippert, X. Wang, F. Ott, J. Muller, C. Alonso-Blanco, K. Borgwardt, K. J. Schmid, and D. Weigel. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nature Genetics*, 43(10):956–963, 2011.

[41] Q. Chen, Z. Han, H. Jiang, D. Tian, and S. Yang. Strong positive selection drives rapid diversification of R-genes in *Arabidopsis* relatives. *J Mol Evol*, 70(2):137–48, 2010.

[42] J. Wang, L. Zhang, J. Li, A. Lawton-Rauh, and D. Tian. Unusual signatures of highly adaptable R-loci in closely-related *Arabidopsis* species. *Gene*, 482(1-2):24–33, 2011.

[43] Shengjun Tan, Yan Zhong, Huan Hou, Sihai Yang, and Dacheng Tian. Variation of presence/absence genes among *Arabidopsis* populations. *BMC Evolutionary Biology*, 12(1):86, 2012.

[44] Toni I. Gossmann, Bao-Hua Song, Aaron J. Windsor, Thomas Mitchell-Olds, Christopher J. Dixon, Maxim V. Kapralov, Dmitry A. Filatov, and Adam Eyre-Walker. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Molecular Biology and Evolution*, 27(8):1822–1832, 2010.

[45] M. Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK, 1983.

[46] Peter D. Keightley, Martin J. Lercher, and Adam Eyre-Walker. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biology*, 3(2):e42, 2005.

[47] S. I. Nikolaev, J. I. Montoya-Burgos, K. Popadin, L. Parand, E. H. Margulies, Program National Institutes of Health Intramural Sequencing Center Comparative Sequencing, and S. E. Antonarakis. Life-history traits drive the evolutionary rates of mammalian coding and noncoding genomic elements. *Proceedings of the National Academy of Sciences of the United States of America*, 104(51):20443–8, 2007.

[48] Konstantin Popadin, Leonard V. Polishchuk, Leila Mamirova, Dmitry Knorre, and Konstantin Gunbin. Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. *Proceedings of the National Academy of Sciences of the United States of America*, 104(33):13390–13395, 2007.

[49] D. Tian, H. Araki, E. Stahl, J. Bergelson, and M. Kreitman. Signature of balancing selection in *Arabidopsis*. *Proceedings of the National Academy of Sciences of the United States of America*, 99(17):11525–11530, 2002.

[50] Peter Szovenyi, Nicolas Devos, David J. Weston, Xiaohan Yang, Zsofia Hock, Jonathan A. Shaw, Kentaro K. Shimizu, Stuart F. McDaniel, and Andreas Wagner. Efficient purging of deleterious mutations in plants with haploid selfing. *Genome Biology and Evolution*, 6(5):1238–1252, 2014.

[51] S. I. Wright, S. Kalisz, and T. Slotte. Evolutionary consequences of self-fertilization in plants. *Proceedings of the Royal Society B: Biological Sciences*, 280(1760):20130133, 2013.

[52] D. Charlesworth and J.H. Willis. The genetics of inbreeding depression. *Nature Reviews Genetics*, 10(11):783–796, 2009.

[53] C.D. Bustamante, R. Nielsen, S.A. Sawyer, K.M. Olsen, M.D. Purugganan, and D.L. Hartl. The cost of inbreeding in *Arabidopsis*. *Nature*, 416:531–534, 2002.

[54] S. Glemin and A. Muyle. Mating systems and selection efficacy: a test using chloroplastic sequence data in angiosperms. *Journal of Evolutionary Biology*, 2014.

[55] Suo Qiu, Kai Zeng, Tanja Slotte, Stephen Wright, and Deborah Charlesworth. Reduced efficacy of natural selection on codon usage bias in selfing *Arabidopsis* and *Capsella* species. *Genome Biology and Evolution*, 3:868–880, 2011.

[56] J. L. Parmley and L. D. Hurst. Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. *Molecular Biology and Evolution*, 24(8):1600–1603, 2007.

[57] T. Warnecke and L. D. Hurst. Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in *Drosophila melanogaster*. *Molecular Biology and Evolution*, 24(12):2755–2762, 2007.

[58] D. B. Carlini and J. E. Genut. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. *J Mol Evol*, 62(1):89–98, 2006.

[59] Joanna L. Parmley, J. V. Chamary, and Laurence D. Hurst. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Molecular Biology and Evolution*, 23(2):301–309, 2006.

[60] M. T. Hamblin, A. M. Casa, H. Sun, S. C. Murray, A. H. Paterson, C. F. Aquadro, and S. Kresovich. Challenges of detecting directional selection after a bottleneck: lessons from *Sorghum bicolor*. *Genetics*, 173(2):953–64, 2006.

[61] Jeffrey Ross-Ibarra, Maud Tenaillon, and Brandon S. Gaut. Historical divergence and gene flow in the genus *Zea*. *Genetics*, 181(4):1399–1413, 2009.

[62] M. Schmid, T. S. Davison, S. R. Henz, U. J. Pape, M. Demar, M. Vingron, B. Scholkopf, D. Weigel, and J. U. Lohmann. A gene expression map of *Arabidopsis thaliana* development. *Nature Genetics*, 37(5):501–506, 2005.

[63] Tanja Slotte, Thomas Bataillon, Troels T. Hansen, Kate St. Onge, Stephen I. Wright, and Mikkel H. Schierup. Genomic determinants of protein evolution and polymorphism in *Arabidopsis*. *Genome Biology and Evolution*, 3:1210–1219, 2011.

[64] John Paul Foxe, Vaqaar-un-Nisa Dar, Honggang Zheng, Magnus Nordborg, Brandon S. Gaut, and Stephen I. Wright. Selection on amino acid substitutions in *Arabidopsis*. *Molecular Biology and Evolution*, 25(7):1375–1383, 2008.

[65] R. A. Swanson-Wagner, S. R. Eichten, S. Kumari, P. Tiffin, J. C. Stein, D. Ware, and N. M. Springer. Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Research*, 20(12):1689–1699, 2010.

[66] Lei-Ying Zheng, Xiao-Sen Guo, Bing He, Lian-Jun Sun, Yao Peng, Shan-Shan Dong, Teng-Fei Liu, Shuye Jiang, Srinivasan Ramachandran, Chun-Ming Liu, and Hai-Chun Jing. Genome-wide patterns of genetic variation in sweet and grain sorghum (*Sorghum bicolor*). *Genome Biology*, 12(11):R114, 2011.

[67] L. K. McHale, W. J. Haun, W. W. Xu, P. B. Bhaskar, J. E. Anderson, D. L. Hyten, D. J. Gerhardt, J. A. Jeddeloh, and R. M. Stupar. Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant Physiology*, 159(4):1295–308, 2012.

[68] Leah McHale, Xiaoping Tan, Patrice Koehl, and Richard Michelmore. Plant NBS-LRR proteins: adaptable guards. *Genome Biology*, 7(4):212, 2006.

[69] Lu Chen, Jaime M. Tovar-Corona, and Araxi O. Urrutia. Alternative splicing: A potential source of functional innovation in the eukaryotic genome. *Int Journal Evol Biol*, 2012:10, 2012.

[70] Brenton R. Graveley. Alternative splicing: increasing diversity in the proteomic world. *Trends in Genetics*, 17(2):100–107, 2001.

[71] Timothy W. Nilsen and Brenton R. Graveley. Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280):457–463, 2010.

[72] Michael Lynch. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proceedings of the National Academy of Sciences of the United States of America*, 104(Suppl 1):8597–8604, 2007.

[73] Michael Lynch and John S. Conery. The origins of genome complexity. *Science*, 302(5649):1401–1404, 2003.

[74] Guido Leoni, Loredana Le Pera, Fabrizio Ferre, Domenico Raimondo, and Anna Tramontano. Coding potential of the products of alternative splicing in human. *Genome Biology*, 12(1):R9, 2011.

[75] E. Melamud and J. Moult. Stochastic noise in splicing machinery. *Nucleic Acids Research*, 37(14):4873–4886, 2009.

[76] J. K. Pickrell, A. A. Pai, Y. Gilad, and J. K. Pritchard. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genetics*, 6(12):e1001236, 2010.

[77] R. Sorek, R. Shamir, and G. Ast. How prevalent is functional alternative splicing in the human genome? *Trends in Genetics*, 20(2):68–71, 2004.

[78] S. Stamm, S. Ben-Ari, I. Rafalska, Y. Tang, Z. Zhang, D. Toiber, T. A. Thanaraj, and H. Soreq. Function of alternative splicing. *Gene*, 344:1–20, 2005.

[79] A. Resch, Y. Xing, B. Modrek, M. Gorlick, R. Riley, and C. Lee. Assessing the impact of alternative splicing on domain interactions in the human proteome. *Journal of Proteome Research*, 3(1):76–83, 2004.

[80] Y. Xing, Q. Xu, and C. Lee. Widespread production of novel soluble protein isoforms by alternative splicing removal of transmembrane anchoring domains. *FEBS Letters*, 555(3):572–8, 2003.

[81] A. D. Neverov, II Artamonova, R. N. Nurtdinov, D. Frishman, M. S. Gelfand, and A. A. Mironov. Alternative splicing and protein function. *BMC Bioinformatics*, 6:266, 2005.

[82] K. Yura, M. Shionyu, K. Hagino, A. Hijikata, Y. Hirashima, T. Nakahara, T. Eguchi, K. Shinoda, A. Yamaguchi, K. Takahashi, T. Itoh, T. Imanishi, T. Gojobori, and M. Go. Alternative splicing in human transcriptome: functional and structural influence on proteins. *Gene*, 380(2):63–71, 2006.

[83] Konrad Grutzmann, Karol Szafranski, Martin Pohl, Kerstin Voigt, Andreas Petzold, and Stefan Schuster. Fungal alternative splicing is associated with multicellular complexity and virulence: A genome-wide multi-species study. *DNA Research*, 21(1):27–39, 2014.

[84] J. A. Calarco, Y. Xing, M. Caceres, J. P. Calarco, X. Xiao, Q. Pan, C. Lee, T. M. Preuss, and B. J. Blencowe. Global analysis of alternative splicing differences between humans and chimpanzees. *Genes and Development*, 21(22):2963–2975, 2007.

[85] L. Lin, S. Shen, P. Jiang, S. Sato, B. L. Davidson, and Y. Xing. Evolution of alternative splicing in primate brain transcriptomes. *Human Molecular Genetics*, 19(15):2958–2973, 2010.

[86] H. Lee, C. Dean, and E. Isacoff. Alternative splicing of neuroligin regulates the rate of presynaptic differentiation. *Journal of Neuroscience*, 30(34):11435–46, 2010.

[87] Fiona L. Watson, Roland Püttmann-Holgado, Franziska Thomas, David L. Lamar, Michael Hughes, Masahiro Kondo, Vivienne I. Rebel, and Dietmar Schmucker. Extensive diversity of Ig-superfamily proteins in the immune system of insects. *Science*, 309(5742):1874–1878, 2005.

[88] Joachim Kurtz and Sophie A.O. Armitage. Alternative adaptive immunity in invertebrates. *Trends in Immunology*, 27(11):493 – 496, 2006.

[89] Helen K Salz. Sex determination in insects: a binary decision based on alternative splicing. *Current Opinion in Genetics & Development*, 21(4):395 – 400, 2011. Differentiation and gene regulation.

[90] W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S, 4th Edition*. Springer, New York, 2002.

[91] Roberto Bonasio, Qiye Li, Jinmin Lian, Navdeep S. Mutti, Lijun Jin, Hongmei Zhao, Pei Zhang, Ping Wen, Hui Xiang, Yun Ding, Zonghui Jin, Steven S. Shen, Zongji Wang, Wen Wang, Jun Wang, Shelley L. Berger, Jürgen Liebig, Guojie Zhang, and Danny Reinberg. Genome-wide and caste-specific DNA methylomes

of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Current Biology*, 22(19):1755 – 1764, 2012.

[92] Frank Lyko, Sylvain Foret, Robert Kucharski, Stephan Wolf, Cassandra Falckenhayn, and Ryszard Maleszka. The honey bee epigenomes: Differential methylation of brain DNA in queens and workers. *PLoS Biology*, 8(11):e1000506, 11 2010.

[93] N. L. Barbosa-Morais, M. Irimia, Q. Pan, H. Y. Xiong, S. Gueroussov, L. J. Lee, V. Slobodeniuc, C. Kutter, S. Watt, R. Colak, T. Kim, C. M. Misquitta-Ali, M. D. Wilson, P. M. Kim, D. T. Odom, B. J. Frey, and B. J. Blencowe. The evolutionary landscape of alternative splicing in vertebrate species. *Science*, 338(6114):1587–1593, 2012.

[94] Gene W. Yeo, Eric Van Nostrand, Dirk Holste, Tomaso Poggio, and Christopher B. Burge. Identification and analysis of alternative splicing events conserved in human and mouse. *Proceedings of the National Academy of Sciences of the United States of America*, 102(8):2850–2855, 2005.

[95] T. A. Thanaraj, Francis Clark, and Juha Muilu. Conservation of human alternative splice events in mouse. *Nucleic Acids Research*, 31(10):2544–2552, 2003.

[96] Ramil Nurtdinov, Alexey Neverov, Alexander Favorov, Andrey Mironov, and Mikhail Gelfand. Conserved and species-specific alternative splicing in mammalian genomes. *BMC Evolutionary Biology*, 7(1):249, 2007.

[97] Alejandro Reyes, Simon Anders, Robert J. Weatheritt, Toby J. Gibson, Lars M. Steinmetz, and Wolfgang Huber. Drift and conservation of differential exon usage across tissues in primate species. *Proceedings of the National Academy of Sciences of the United States of America*, 2013.

[98] Eddo Kim, Alon Magen, and Gil Ast. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Research*, 35(1):125–131, 2007.

[99] Eva Schad, Peter Tompa, and Hedi Hegyi. The relationship between proteome size, structural disorder and organism complexity. *Genome Biology*, 12(12):R120, 2011.

[100] Mihaela Pertea, Stephen Mount, and Steven Salzberg. A computational survey of candidate exonic splicing enhancer motifs in the model plant *Arabidopsis thaliana*. *BMC Bioinformatics*, 8(1):159, 2007.

[101] Lindell Bromham. Why do species vary in their rate of molecular evolution? *Biology Letters*, 5(3):401–404, 2009.

[102] Laurence D. Hurst. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends in Genetics*, 18(9):486, 2002.

[103] N C Kyrpides. Genomes OnLine database (GOLD 1.0): a monitor of complete and ongoing genome projects world-wide. *Bioinformatics*, 15(9):773–774, 1999.

[104] I. Pagani, K. Liolios, J. Jansson, I. M. Chen, T. Smirnova, B. Nosrat, V. M. Markowitz, and N. C. Kyrpides. The Genomes OnLine database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research*, 40(Database issue):D571–D579, 2012.

[105] T. T. Hu, P. Pattyn, E. G. Bakker, J. Cao, J. F. Cheng, R. M. Clark, N. Fahlgren, J. A. Fawcett, J. Grimwood, H. Gundlach, G. Haberer, J. D. Hollister, S. Ossowski, R. P. Ottilar, A. A. Salamov, K. Schneeberger, M. Spannagl, X. Wang, L. Yang, M. E. Nasrallah, J. Bergelson, J. C. Carrington, B. S. Gaut, J. Schmutz, K. F. Mayer, Y. Van de Peer, I. V. Grigoriev, M. Nordborg, D. Weigel, and Y. L. Guo. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nature Genetics*, 43(5):476–481, 2011.

[106] M. A. Beilstein, N. S. Nagalingum, M. D. Clements, S. R. Manchester, and S. Mathews. Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis* thaliana. *Proceedings of the National Academy of Sciences*, 107(43):18724–18728, 2010.

[107] M. Dassanayake, D. H. Oh, J. S. Haas, A. Hernandez, H. Hong, S. Ali, D. J. Yun, R. A. Bressan, J. K. Zhu, H. J. Bohnert, and J. M. Cheeseman. The genome of the extremophile crucifer *Thellungiella parvula*. *Nat Genet*, 43(9):913–8, 2011.

[108] Liang Yang and Brandon S. Gaut. Factors that contribute to variation in evolutionary rate among *Arabidopsis* genes. *Molecular Biology and Evolution*, 28(8):2359–2369, 2011.

[109] Yongchun Wu, Yongqing Zhang, and Jiong Zhang. Distribution of exonic splicing enhancer elements in human genes. *Genomics*, 86(3):329 – 336, 2005.

[110] Tobias Warnecke, Joanna Parmley, and Laurence Hurst. Finding exonic islands in a sea of non-coding sequence: splicing related constraints on protein composition and evolution are common in intron-rich genomes. *Genome Biology*, 9(2):R29, 2008.

[111] H. Akashi. Translational selection and yeast proteome evolution. *Genetics*, 164(4):1291–1303, 2003.

[112] J. L. Cherry. Expression level, evolutionary rate, and the cost of expression. *Genome Biol Evol*, 2:757–69, 2010.

[113] D. Allan Drummond, Alpan Raval, and Claus O. Wilke. A single determinant dominates the rate of yeast protein evolution. *Molecular Biology and Evolution*, 23(2):327–337, 2006.

[114] D. M. Krylov, Y. I. Wolf, I. B. Rogozin, and E. V. Koonin. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Research*, 13(10):2229–2235, 2003.

[115] B. Y. Liao and J. Zhang. Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution. *Molecular Biology and Evolution*, 23(6):1119–1128, 2006.

[116] T. Paape, T. Bataillon, P. Zhou, J. Y. Kono T, R. Briskine, N. D. Young, and P. Tiffin. Selection, genome-wide fitness effects and evolutionary rates in the model legume *Medicago truncatula*. *Molecular Ecology*, 22(13):3525–3538, 2013.

[117] I. Pagan, E. C. Holmes, and E. Simon-Loriere. Level of gene expression is a major determinant of protein evolution in the viral order mononegavirales. *Journal of Virology*, 86(9):5253–5263, 2012.

[118] C. Pal, B. Papp, and M. J. Lercher. An integrated view of protein evolution. *Nature Reviews. Genetics*, 7(5):337–348, 2006.

[119] E. P. Rocha and A. Danchin. An analysis of determinants of amino acids substitution rates in bacterial proteins. *Molecular Biology and Evolution*, 21(1):108–116, 2004.

[120] P. M. Sharp. Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: codon usage, map position, and concerted evolution. *Journal of Molecular Evolution*, 33(1):23–33, 1991.

[121] Sankar Subramanian and Sudhir Kumar. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics*, 168(1):373–381, 2004.

[122] Laurent Duret and Dominique Mouchiroud. Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate. *Molecular Biology and Evolution*, 17(1):68–70, 2000.

[123] Seung Park and Sun Choi. Expression breadth and expression abundance behave differently in correlations with evolutionary rates. *BMC Evolutionary Biology*, 10(1):241, 2010.

[124] Eitan E. Winter, Leo Goodstadt, and Chris P. Ponting. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Research*, 14(1):54–61, 2004.

[125] Liqing Zhang and Wen-Hsiung Li. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Molecular Biology and Evolution*, 21(2):236–239, 2004.

[126] Araxi O. Urrutia and Laurence D. Hurst. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics*, 159(3):1191–1199, 2001.

[127] Yu Xia, Eric A. Franzosa, and Mark B. Gerstein. Integrated assessment of genomic correlates of protein evolutionary rate. *PLoS Computational Biology*, 5(6):e1000413, 06 2009.

[128] K. Tsunoyama, M. I. Bellgard, and T. Gojobori. Intragenic variation of synonymous substitution rates is caused by nonrandom mutations at methylated CpG. *Journal of Molecular Evolution*, 53(4-5):456–464, 2001.

[129] P. M. Sharp and W. H. Li. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Molecular Biology and Evolution*, 4(3):222–230, 1987.

[130] A. Eyre-Walker and M. Bulmer. Synonymous substitution rates in enterobacteria. *Genetics*, 140(4):1407–1412, 1995.

[131] Hiroshi Akashi. Gene expression and molecular evolution. *Current Opinion in Genetics & Development*, 11(6):660–666, 2001.

[132] A. Ticher and D. Grauer. Nucleic acid composition, codon usage, and the rate of synonymous substitution in protein-coding genes. *Journal of Molecular Evolution*, 28(4):286–298, 1989.

[133] J.L. Cherry. Highly expressed and slowly evolving proteins share compositional properties with thermophilic proteins. *Molecular Biology and Evolution*, 27(3):735–741, 2010.

[134] Matthew W. Hahn and Andrew D. Kern. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular Biology and Evolution*, 22(4):803–806, 2005.

[135] Soumita Podder, Pamela Mukhopadhyay, and Tapash Chandra Ghosh. Multifunctionality dominantly determines the rate of human housekeeping and tissue specific interacting protein evolution. *Gene*, 439(1-2):11–16, 2009.

[136] Bratati Kahali, Shandar Ahmad, and Tapash Chandra Ghosh. Exploring the evolutionary rate differences of party hub and date hub proteins in *Saccharomyces cerevisiae* protein-protein interaction network. *Gene*, 429(1-2):18–22, 2009.

[137] Hunter B. Fraser. Modularity and evolutionary constraint on proteins. *Nature Genetics*, 37:351–352, 2005.

[138] Takashi Makino and Takashi Gojobori. The evolutionary rate of a protein is influenced by features of the interacting partners. *Molecular Biology and Evolution*, 23(4):784–789, 2006.

[139] Guang-Zhong Wang and Martin J. Lercher. The effects of network neighbours on protein evolution. *PLoS ONE*, 6(4):e18288, 04 2011.

[140] Hunter B. Fraser and Aaron E. Hirsh. Evolutionary rate depends on number of protein-protein interactions independently of gene expression level. *BMC Evolutionary Biology*, 4(1), 2004.

[141] Csaba Pal, Balazs Papp, and Laurence D. Hurst. Does the recombination rate affect the efficiency of purifying selection? the yeast genome provides a partial answer. *Molecular Biology and Evolution*, 18(12):2323–2326, 2001.

[142] Stephen I. Wright, John Paul Foxe, Leah DeRose-Wilson, Akira Kawabe, Mark Looseley, Brandon S. Gaut, and Deborah Charlesworth. Testing for effects of recombination rate on nucleotide diversity in natural populations of *Arabidopsis lyrata*. *Genetics*, 174(3):1421–1430, 2006.

[143] A. O. Urrutia and L. D. Hurst. The signature of selection mediated by expression on human genes. *Genome Research*, 13(10):2260–2264, 2003.

[144] B. Lemos, B. R. Bettencourt, C. D. Meiklejohn, and D. L. Hartl. Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. *Molecular Biology and Evolution*, 22(5):1345–54, 2005.

[145] A. Coghlan and K. H. Wolfe. Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. *Yeast*, 16(12):1131–45, 2000.

[146] Amanda M. Larracuente, Timothy B. Sackton, Anthony J. Greenberg, Alex Wong, Nadia D. Singh, David Sturgill, Yu Zhang, Brian Oliver, and Andrew G. Clark. Evolution of protein-coding genes in *Drosophila*. *Trends in Genetics*, 24(3):114–123, 2008.

[147] Cathal Seoighe, Chris Gehring, and Laurence D. Hurst. Gametophytic selection in *Arabidopsis thaliana* supports the selective model of intron length reduction. *PLoS Genetics*, 1(2):e13, 08 2005.

[148] Clara S. Tang, Yong Z. Zhao, David K. Smith, and Richard J. Epstein. Intron length and accelerated 3' gene evolution. *Genomics*, 88(6):682–689, 2006.

[149] I. Yanai, H. Benjamin, M. Shmoish, V. Chalifa-Caspi, M. Shklar, R. Ophir, A. Bar-Even, S. Horn-Saban, M. Safran, E. Domany, D. Lancet, and O. Shmueli. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, 21(5):650–9, 2005.

[150] William Revelle. psych: Procedures for psychological, psychometric, and personality research, 2014.

[151] William G. Fairbrother, Dirk Holste, Christopher B. Burge, and Phillip A. Sharp. Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS Biology*, 2(9):e268, 2004.

[152] Lu Chen, Stephen J. Bush, Jaime M. Tovar-Corona, Atahualpa Castillo-Morales, and Araxi O. Urrutia. Correcting for differential transcript coverage reveals a strong relationship between alternative splicing and organism complexity. *Molecular Biology and Evolution*, 31(6):1402–1413, 2014.

[153] Tomasz E. Koralewski and Konstantin V. Krutovsky. Evolution of exon-intron structure and alternative splicing. *PLoS ONE*, 6(3):e18055, 2011.

[154] Liucun Zhu, Ying Zhang, Wen Zhang, Sihai Yang, Jian-Qun Chen, and Dacheng Tian. Patterns of exon-intron architecture variation of genes in eukaryotic genomes. *BMC Genomics*, 10(1):47, 2009.

[155] Sylvain Glemin. Mating systems and the efficacy of selection at the molecular level. *Genetics*, 177(2):905–916, 2007.

[156] G. D. Morrison and C. R. Linder. Association mapping of germination traits in *Arabidopsis thaliana* under light and nutrient treatments: Searching for G x E effects. *G3 (Bethesda)*, 2014.

[157] Sanjay Singh, Sujit Roy, Swarup Choudhury, and Dibyendu Sengupta. DNA repair and recombination in higher plants: insights from comparative genomics of *Arabidopsis* and rice. *BMC Genomics*, 11(1):443, 2010.

[158] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[159] W. R. Pearson. Flexible sequence similarity searching with the FASTA3 program package. *Methods in Molecular Biology*, 132:185–219, 2000.

[160] Ziheng Yang. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591, 2007.

[161] Ari Loytynoja and Nick Goldman. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, 320(5883):1632–1635, 2008.

[162] J. B. Haldane. The estimation and significance of the logarithm of a ratio of frequencies. *Annals of Human Genetics*, 20(4):309–11, 1956.

[163] J. H. McDonald and M. Kreitman. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature*, 351(6328):652–654, 1991.

[164] Mark S. Boguski, Todd M. J. Lowe, and Carolyn M. Tolstoshev. dbEST - database for expressed sequence tags. *Nature Genetics*, 4(4):332–333, 1993.

[165] Lu Chen, Jaime M. Tovar-Corona, and Araxi O. Urrutia. Increased levels of noisy splicing in cancers, but not for oncogene-derived transcripts. *Human Molecular Genetics*, 20(22):4422–4429, 2011.

[166] Zhijin Wu, Rafael A Irizarry, Robert Gentleman, Francisco Martinez-Murillo, and Forrest Spencer. A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association*, 99(468):909–917, 2004.

[167] S. Brenner, M. Johnson, J. Bridgham, G. Golda, D. H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M. Ewan, R. Roth, D. George, S. Eletr, G. Albrecht, E. Vermaas, S. R. Williams, K. Moon, T. Burcham, M. Pallas, R. B. DuBridge, J. Kirchner, K. Fearon, J. Mao, and K. Corcoran. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnology*, 18(6):630–634, 2000.

[168] B. C. Meyers, S. S. Tej, T. H. Vu, C. D. Haudenschild, V. Agrawal, S. B. Edberg, H. Ghazal, and S. Decola. The use of MPSS for whole-genome transcriptional analysis in *Arabidopsis*. *Genome Research*, 14(8):1641–1653, 2004.

[169] Mayumi Nakano, Kan Nobuta, Kalyan Vemaraju, Shivakundan Singh Tej, Jeremy W. Skogen, and Blake C. Meyers. Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. *Nucleic Acids Research*, 34(suppl 1):D731–D735, 2006.

[170] Hangxing Yang. In plants, expression breadth and expression level distinctly and non-linearly correlate with gene structure. *Biology Direct*, 4(1):45, 2009.

[171] C. Cheadle, M. P. Vawter, W. J. Freed, and K. G. Becker. Analysis of microarray data using Z score transformation. *Journal of Molecular Diagnostics*, 5(2):73–81, 2003.

[172] Katja Baerenfaller, Jonas Grossmann, Monica A. Grobei, Roger Hull, Matthias Hirsch-Hoffmann, Shaul Yalovsky, Philip Zimmermann, Ueli Grossniklaus, Wilhelm Gruissem, and Sacha Baginsky. Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. *Science*, 320(5878):938–941, 2008.

[173] Natalie E. Castellana, Samuel H. Payne, Zhouxin Shen, Mario Stanke, Vineet Bafna, and Steven P. Briggs. Discovery and revision of *Arabidopsis* genes by proteogenomics. *Proceedings of the National Academy of Sciences*, 105(52):21034–21038, 2008.

[174] F. Wright. The 'effective number of codons' used in a gene. *Gene*, 87:23–29, 1990.

[175] T. Ikemura. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *Journal of Molecular Biology*, 151:389–409, 1981.

[176] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, 34(Database issue):D535–539, 2006.

[177] Chris Stark, Bobby-Joe Breitkreutz, Andrew Chatr-aryamontri, Lorrie Boucher, Rose Oughtred, Michael S. Livstone, Julie Nixon, Kimberly Van Auken, Xiaodong Wang, Xiaoqi Shi, Teresa Reguly, Jennifer M. Rust, Andrew Winter, Kara Dolinski, and Mike Tyers. The BioGRID interaction database: 2011 update. *Nucleic Acids Research*, 39(suppl 1):D698–D704, 2011.

[178] G. Marais, B. Charlesworth, and S. I. Wright. Recombination and base composition: the case of the highly self-fertilizing plant *Arabidopsis thaliana*. *Genome Biology*, 5:R45, 2004.

[179] Lars Feuk, Andrew R. Carson, and Stephen W. Scherer. Structural variation in the human genome. *Nature Reviews. Genetics*, 7(2):85–97, 2006.

[180] Yi Wang, Frank M. You, Gerard R. Lazo, Ming-Cheng Luo, Roger Thilmony, Sean Gordon, Shahryar F. Kianian, and Yong Q. Gu. PIECE: a database for plant gene structure comparison and evolution. *Nucleic Acids Research*, 41(D1):D1159–D1166, 2013.

[181] Xun Xu, Xin Liu, Song Ge, Jeffrey D. Jensen, Fengyi Hu, Xin Li, Yang Dong, Ryan N. Gutenkunst, Lin Fang, Lei Huang, Jingxiang Li, Weiming He, Guojie Zhang, Xiaoming Zheng, Fumin Zhang, Yingrui Li, Chang Yu, Karsten Kristiansen, Xiuqing Zhang, Jian Wang, Mark Wright, Susan McCouch, Rasmus Nielsen, Jun Wang, and Wen Wang. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nature Biotechnology*, 30(1):105–111, 2012.

[182] Richard M. Clark, Gabriele Schweikert, Christopher Toomajian, Stephan Ossowski, Georg Zeller, Paul Shinn, Norman Warthmann, Tina T. Hu, Glenn Fu,

David A. Hinds, Huaming Chen, Kelly A. Frazer, Daniel H. Huson, Bernhard Scholkopf, Magnus Nordborg, Gunnar Ratsch, Joseph R. Ecker, and Detlef Weigel. Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science*, 317(5836):338–342, 2007.

[183] S. Ossowski, K. Schneeberger, R. M. Clark, C. Lanz, N. Warthmann, and D. Weigel. Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Research*, 18(12):2024–2033, 2008.

[184] E. G. Bakker, C. Toomajian, M. Kreitman, and J. Bergelson. A genome-wide survey of R gene polymorphisms in *Arabidopsis*. *Plant Cell*, 18(8):1803–1818, 2006.

[185] J. Shen, H. Araki, L. Chen, J. Q. Chen, and D. Tian. Unique evolutionary mechanism in R-genes under the presence/absence polymorphism in *Arabidopsis thaliana*. *Genetics*, 172(2):1243–1250, 2006.

[186] Luca Santuari, Sylvain Pradervand, Amelia-Maria Amiguet-Vercher, Jerome Thomas, Eavan Dorcey, Keith Harshman, Ioannis Xenarios, Thomas Juenger, and Christian Hardtke. Substantial deletion overlap among divergent *Arabidopsis* genomes revealed by intersection of short reads and tiling arrays. *Genome Biology*, 11(1):R4, 2010.

[187] Brody J. DeYoung and Roger W. Innes. Plant NBS-LRR proteins in pathogen sensing and host defense. *Nature Immunology*, 7(12):1243–1249, 2006.

[188] K. Hanada, T. Kuromori, F. Myouga, T. Toyoda, W. H. Li, and K. Shinozaki. Evolutionary persistence of functional compensation by duplicate genes in *Arabidopsis*. *Genome Biology and Evolution*, 1:409–14, 2009.

[189] T. Makino, K. Hokamp, and A. McLysaght. The complex relationship of gene duplication and essentiality. *Trends in Genetics*, 25(4):152–155, 2009.

[190] Yuri I Wolf, Liran Carmel, and Eugene V Koonin. Unifying measures of gene function and evolution. *Proceedings of the Royal Society B: Biological Sciences*, 273(1593):1507–1515, 2006.

[191] S. I. Wright, N. Agrawal, and T. E. Bureau. Effects of recombination rate and gene density on transposable element distributions in *Arabidopsis thaliana*. *Genome Research*, 13(8):1897–1903, 2003.

[192] Matthew W. Horton, Angela M. Hancock, Yu S. Huang, Christopher Toomajian, Susanna Atwell, Adam Auton, N. Wayan Muliyati, Alexander Platt, F. Gianluca Sperone, Bjarni J. Vilhjalmsson, Magnus Nordborg, Justin O. Borevitz, and Joy Bergelson. Genome-wide patterns of genetic variation in worldwide *Arabidopsis*

thaliana accessions from the RegMap panel. *Nature Genetics*, 44(2):212–216, 2012.

[193] K. R. Oliver and W. K. Greene. Transposable elements: powerful facilitators of evolution. *Bioessays*, 31(7):703–714, 2009.

[194] M. Mondragon-Palomino, B. C. Meyers, R. W. Michelmore, and B. S. Gaut. Patterns of positive selection in the complete NBS-LRR gene family of *Arabidopsis thaliana*. *Genome Research*, 12(9):1305–1315, 2002.

[195] Koichiro Tamura and Sudhir Kumar. Evolutionary distance estimation under heterogeneous substitution pattern among lineages. *Molecular Biology and Evolution*, 19(10):1727–1736, 2002.

[196] Angela M. Hancock, Benjamin Brachi, Nathalie Faure, Matthew W. Horton, Lucien B. Jarymowycz, F. Gianluca Sperone, Chris Toomajian, Fabrice Roux, and Joy Bergelson. Adaptation to climate across the *Arabidopsis thaliana* genome. *Science*, 334(6052):83–86, 2011.

[197] Christian D. Huber, Magnus Nordborg, Joachim Hermisson, and Ines Hellmann. Keeping it local: Evidence for positive selection in Swedish *Arabidopsis thaliana*. *Molecular Biology and Evolution*, 2014.

[198] Renier A. L. Van der Hoorn, Pierre J. G. M. De Wit, and Matthieu H. A. J. Joosten. Balancing selection favors guarding resistance proteins. *Trends in Plant Science*, 7(2):67–71, 2002.

[199] G. Gos and S. I. Wright. Conditional neutrality at two adjacent NBS-LRR disease resistance loci in natural populations of *Arabidopsis lyrata*. *Molecular Ecology*, 17(23):4953–62, 2008.

[200] F Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595, 1989.

[201] Barmak Modrek and Christopher J. Lee. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nature Genetics*, 34(2):177–180, 2003.

[202] Xi Wang, Detlef Weigel, and Lisa M. Smith. Transposon variants and their effects on gene expression in *Arabidopsis*. *PLoS Genetics*, 9(2):e1003255, 2013.

[203] H. R Johnston and D. J Cutler. Population demographic history can cause the appearance of recombination hotspots. *The American Journal of Human Genetics*, 90(5):774–783, 2012.

[204] Casey M. Bergman and Hadi Quesneville. Discovering and detecting transposable elements in genome sequences. *Briefings in Bioinformatics*, 8(6):382–392, 2007.

[205] Damian Smedley, Syed Haider, Benoit Ballester, Richard Holland, Darin London, Gudmundur Thorisson, and Arek Kasprzyk. BioMart - biological queries made easy. *BMC Genomics*, 10(1):22, 2009.

[206] Marco Punta, Penny C. Coggill, Ruth Y. Eberhardt, Jaina Mistry, John Tate, Chris Boursnell, Ningze Pang, Kristoffer Forslund, Goran Ceric, Jody Clements, Andreas Heger, Liisa Holm, Erik L. L. Sonnhammer, Sean R. Eddy, Alex Bateman, and Robert D. Finn. The Pfam protein families database. *Nucleic Acids Research*, 40(D1):D290–D301, 2012.

[207] A. J. Vilella, J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, and E. Birney. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research*, 19(2):327–335, 2009.

[208] Tomislav Domazet-Loso, Josip Brajkovic, and Diethard Tautz. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends in Genetics*, 23(11):533–539, 2007.

[209] H. Kuittinen, A. A. de Haan, C. Vogl, S. Oikarinen, J. Leppala, M. Koch, T. Mitchell-Olds, C. H. Langley, and O. Savolainen. Comparing the linkage maps of the close relatives *Arabidopsis lyrata* and *A. thaliana*. *Genetics*, 168(3):1575–1584, 2004.

[210] Ryan Taft and John Mattick. Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding DNA sequences. *Genome Biology*, 5(1):P1, 2003.

[211] A. P. Bird. Gene number, noise reduction and biological complexity. *Trends in Genetics*, 11(3):94–100, 1995.

[212] C. Fields, M. D. Adams, O. White, and J. C. Venter. How many genes in the human genome? *Nature Genetics*, 7(3):345–346, 1994.

[213] P. Dehal and J. L. Boore. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol*, 3(10):e314, 2005.

[214] Susumu Ohno. *Evolution by gene duplication*. Springer-Verlag, New York, 1970.

[215] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas,

A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[216] Kai Xia, Zheng Fu, Lei Hou, and Jing-Dong J. Han. Impacts of protein-protein interaction domains on organism and network complexity. *Genome Research*, 18(9):1500–1508, 2008.

[217] A. K. Dunker, C. J. Oldfield, J. Meng, P. Romero, J. Y. Yang, J. W. Chen, V. Vacic, Z. Obradovic, and V. N. Uversky. The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics*, 9 Suppl 2:S1, 2008.

[218] Pedro R. Romero, Saima Zaidi, Ya Yin Fang, Vladimir N. Uversky, Predrag Radivojac, Christopher J. Oldfield, Marc S. Cortese, Megan Sickmeier, Tanguy LeGall, Zoran Obradovic, and A. Keith Dunker. Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proceedings of the National Academy of Sciences of the United States of America*, 103(22):8390–8395, 2006.

[219] Y. Xing and C. Lee. Relating alternative splicing to proteome complexity and genome evolution. *Advances in Experimental Medicine and Biology*, 623:36–49, 2007.

[220] Qun Pan, Ofer Shai, Leo J. Lee, Brendan J. Frey, and Benjamin J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40(12):1413–1415, 2008.

[221] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 2008.

[222] Heebal Kim, Robert Klein, Jacek Majewski, and Jurg Ott. Estimating rates of alternative splicing in mammals and invertebrates. *Nature Genetics*, 36(9):915–916, 2004.

[223] D. Brett, H. Pospisil, J. Valcarcel, J. Reich, and P. Bork. Alternative splicing and genome complexity. *Nature Genetics*, 30(1):29–30, 2002.

[224] I. G. Mollet, Claudia Ben-Dov, Daniel Felicio-Silva, A. R. Grosso, Pedro Eleuterio, Ruben Alves, Ray Staller, Tito Santos Silva, and Maria Carmo-Fonseca. Unconstrained mining of transcript data reveals increased alternative splicing complexity in the human transcriptome. *Nucleic Acids Research*, 38(14):4740–4754, 2010.

[225] Jun-ichi Takeda, Yutaka Suzuki, Ryuichi Sakate, Yoshiharu Sato, Masahide Seki, Takuma Irie, Nono Takeuchi, Takuya Ueda, Mitsuteru Nakao, Sumio Sugano, Takashi Gojobori, and Tadashi Imanishi. Low conservation and species-specific evolution of alternative splicing in humans and mice: comparative genomics analysis using well-annotated full-length cDNAs. *Nucleic Acids Research*, 36(20):6386–6395, 2008.

[226] Paul M. Harrison, Anuj Kumar, Ning Lang, Michael Snyder, and Mark Gerstein. A question of size: the eukaryotic proteome and the problems in defining it. *Nucleic Acids Research*, 30(5):1083–1090, 2002.

[227] Maria Warnefors and Adam Eyre-Walker. The accumulation of gene regulation through time. *Genome Biology and Evolution*, 3:667–673, 2011.

[228] B. Xue, A. K. Dunker, and V. N. Uversky. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *Journal of Biomolecular Structure and Dynamics*, 30(2):137–49, 2012.

[229] Philippe Gayral, Jose Melo-Ferreira, Sylvain Glemin, Nicolas Bierne, Miguel Carneiro, Benoit Nabholz, Joao M. Lourenco, Paulo C. Alves, Marion Ballenghien, Nicolas Faivre, Khalid Belkhir, Vincent Cahais, Etienne Loire, Aurelien Bernard, and Nicolas Galtier. Free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap. *PLoS Genetics*, 9(4):e1003457, 2013.

[230] S. Blair Hedges, Joel Dudley, and Sudhir Kumar. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, 22(23):2971–2972, 2006.

[231] F. Delsuc, H. Brinkmann, D. Chourrout, and H. Philippe. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature*, 439(7079):965–8, 2006.

[232] M. Pagel. Inferring the historical patterns of biological evolution. *Nature*, 401(6756):877–884, 1999.

[233] Elsa Chacko and Shoba Ranganathan. Comprehensive splicing graph analysis of alternative splicing patterns in chicken, compared to human and mouse. *BMC Genomics*, 10(Suppl 1):S5, 2009.

[234] Daniel W. McShea. Functional complexity in organisms: Parts as proxies. *Biology and Philosophy*, 15(5):641–668, 2000.

[235] C. Adami. What is complexity? *Bioessays*, 24(12):1085–1094, 2002.

[236] Olivier Tenaillon, Olin K. Silander, Jean-Philippe Uzan, and Lin Chao. Quantifying organismal complexity using a population genetic approach. *PLoS ONE*, 2(2):e217, 2007.

[237] C. H. Chen, H. Y. Lin, C. L. Pan, and F. C. Chen. The plausible reason why the length of 5' untranslated region is unrelated to organismal complexity. *BMC Res Notes*, 4:312, 2011.

[238] E. Betran and M. Long. Expansion of genome coding regions by acquisition of new genes. *Genetica*, 115(1):65–80, 2002.

[239] Matthew W. Hahn and Gregory A. Wray. The G-value paradox. *Evolution and Development*, 4(2):73–75, 2002.

[240] Jean-Michel Claverie. What if there are only 30,000 human genes? *Science*, 291(5507):1255–1257, 2001.

[241] M. Buljan, G. Chalancon, S. Eustermann, G. P. Wagner, M. Fuxreiter, A. Bateman, and M. M. Babu. Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Molecular Cell*, 46(6):871–883, 2012.

[242] Matteo Floris, Massimiliano Orsini, and Thangavel Thanaraj. Splice-mediated Variants of Proteins (SpliVaP) - data and characterization of changes in signatures among protein isoforms due to alternative splicing. *BMC Genomics*, 9(1):453, 2008.

[243] Evgenia V. Kriventseva, Ina Koch, Rolf Apweiler, Martin Vingron, Peer Bork, Mikhail S. Gelfand, and Shamil Sunyaev. Increase of functional diversity by alternative splicing. *Trends in Genetics*, 19(3):124–128, 2003.

[244] Arli A. Parikesit, Peter F. Stadler, and Sonja J. Prohaska. Evolution and quantitative comparison of genome-wide protein domain distributions. *Genes*, 2(4):912–924, 2011.

[245] M. K. Basu, L. Carmel, I. B. Rogozin, and E. V. Koonin. Evolution of protein domain promiscuity in eukaryotes. *Genome Research*, 18(3):449–461, 2008.

[246] C. H. Kuo, N. A. Moran, and H. Ochman. The consequences of genetic drift for bacterial genome complexity. *Genome Research*, 19(8):1450–1454, 2009.

[247] Kenneth D. Whitney and Jr. Garland, Theodore. Did genetic drift drive increases in genome complexity? *PLoS Genetics*, 6(8):e1001080, 2010.

[248] Kenneth D. Whitney, Bastien Boussau, Eric J. Baack, and Jr. Garland, Theodore. Drift and genome complexity revisited. *PLoS Genetics*, 7(6):e1002092, 2011.

[249] S. A. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology*, 22(3):437–467, 1969.

[250] R. Serra, M. Villani, A. Barbieri, S. A. Kauffman, and A. Colacci. On the dynamics of random Boolean networks subject to noise: attractors, ergodic sets and cell types. *Journal of Theoretical Biology*, 265(2):185–193, 2010.

[251] Bjorn Samuelsson and Carl Troein. Superpolynomial growth in the number of attractors in Kauffman networks. *Physical Review Letters*, 90(9):098701, 2003.

[252] A. A. Kanapin, N. Mulder, and V. A. Kuznetsov. Projection of gene-protein networks to the functional space of the proteome and its application to analysis of organism complexity. *BMC Genomics*, 11 Suppl 1:S4, 2010.

[253] James W. Valentine, Allen G. Collins, and C. Porter Meyer. Morphological complexity increase in metazoans. *Paleobiology*, 20(2):131–142, 1994.

[254] D. Lang, B. Weiche, G. Timmerhaus, S. Richardt, D. M. Riano-Pachon, L. G. Correa, R. Reski, B. Mueller-Roeber, and S. A. Rensing. Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion, and correlation with complexity. *Genome Biology and Evolution*, 2:488–503, 2010.

[255] S Hedges, Jaime Blair, Maria Venturi, and Jason Shoe. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC Evolutionary Biology*, 4(1):2, 2004.

[256] Graham Bell and Arne O. Mooers. Size and complexity among multicellular organisms. *Biological Journal of the Linnean Society of London*, 60(3):345–363, 1997.

[257] R. Haygood and Smbe Tri-National Young Investigators. Proceedings of the SMBE tri-national young investigators' workshop 2005. Mutation rate and the cost of complexity. *Molecular Biology and Evolution*, 23(5):957–63, 2006.

[258] D. H. Erwin. Early origin of the bilaterian developmental toolkit. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 364(1527):2253–2261, 2009.

[259] Christine Vogel and Cyrus Chothia. Protein family expansions and biological complexity. *PLoS Computational Biology*, 2(5):e48, 2006.

[260] Matthew K. Vickaryous and Brian K. Hall. Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biological Reviews*, 81(3):425–455, 2006.

[261] T. D. Wu and C. K. Watanabe. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9):1859–1875, 2005.

[262] Rhoda J. Kinsella, Andreas Kahari, Syed Haider, Jorge Zamora, Glenn Proctor, Giulietta Spudich, Jeff Almeida-King, Daniel Staines, Paul Derwent, Arnaud Kerhornou, Paul Kersey, and Paul Flicek. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database*, 2011, 2011.

[263] Nicholas H. Putnam, Thomas Butts, David E. K. Ferrier, Rebecca F. Furlong, Uffe Hellsten, Takeshi Kawashima, Marc Robinson-Rechavi, Eiichi Shoguchi, Astrid Terry, Jr-Kai Yu, Elia Benito-Gutierrez, Inna Dubchak, Jordi Garcia-Fernandez, Jeremy J. Gibson-Brown, Igor V. Grigoriev, Amy C. Horton, Pieter J. de Jong, Jerzy Jurka, Vladimir V. Kapitonov, Yuji Kohara, Yoko Kuroki, Erika Lindquist, Susan Lucas, Kazutoyo Osoegawa, Len A. Pennacchio, Asaf A. Salamov, Yutaka Satou, Tatjana Sauka-Spengler, Jeremy Schmutz, Tadasu Shin-I, Atsushi Toyoda, Marianne Bronner-Fraser, Asao Fujiyama, Linda Z. Holland, Peter W. H. Holland, Nori Satoh, and Daniel S. Rokhsar. The *Amphioxus* genome and the evolution of the chordate karyotype. *Nature*, 453(7198):1064–1071, 2008.

[264] Robert D. Finn, Jody Clements, and Sean R. Eddy. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39(suppl 2):W29–W37, 2011.

[265] Robert D. Finn, John Tate, Jaina Mistry, Penny C. Coggill, Stephen John Sammut, Hans-Rudolf Hotz, Goran Ceric, Kristoffer Forslund, Sean R. Eddy, Erik L. L. Sonnhammer, and Alex Bateman. The Pfam protein families database. *Nucleic Acids Research*, 36(suppl 1):D281–D288, 2008.

[266] R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012.

[267] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

[268] Sebastien Le, Julie Josse, and Francois Husson. FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software*, 25(1), 2008.

[269] Jose Pinheiro, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, and R Development Core Team. nlme: Linear and nonlinear mixed effects models. r package version 3-1.113, 2013.

[270] David Orme, Rob Freckleton, Gavin Thomas, Thomas Petzoldt, Susanne Fritz, Nick Isaac, and Will Pearse. caper: Comparative analyses of phylogenetics and evolution in r. r package version 0.5, 2012.

[271] Joseph Felsenstein. Phylogenies and the comparative method. *The American Naturalist*, 125(1):1–15, 1985.

[272] A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):6062–6067, 2004.

[273] Maria D. Chikina, Curtis Huttenhower, Coleen T. Murphy, and Olga G. Troyanskaya. Global prediction of tissue-specific gene expression and context-dependent gene networks in *Caenorhabditis elegans*. *PLoS Comput Biol*, 5(6):e1000417, 2009.

[274] Francesca Chiaromonte, Webb Miller, and Eric E. Bouhassira. Gene length and proximity to neighbors affect genome-wide expression levels. *Genome Research*, 13(12):2602–2608, 2003.

[275] You Rao, Zhang Wang, Xue Chai, Guo Wu, Ming Zhou, Qing Nie, and Xi Zhang. Selection for the compactness of highly expressed genes in *Gallus gallus*. *Biology Direct*, 5(1):35, 2010.

[276] Xin-Ying Ren, Oscar Vorst, Mark W.E.J. Fiers, Willem J. Stiekema, and Jan-Peter Nap. In plants, highly expressed genes are the least compact. *Trends in Genetics*, 22(10):528 – 532, 2006.

[277] Rasmus Nielsen, Carlos Bustamante, Andrew G Clark, Stephen Glanowski, Timothy B Sackton, Melissa J Hubisz, Adi Fledel-Alon, David M Tanenbaum, Daniel Civello, Thomas J White, John J. Sninsky, Mark D Adams, and Michele Cargill. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biology*, 3(6):e170, 05 2005.

[278] Haruo Suzuki and Michael J. Stanhope. Functional bias of positively selected genes in *Streptococcus* genomes. *Infection, Genetics and Evolution*, 12(2):274 – 277, 2012.

[279] Julien Roux, Eyal Privman, Sebastien Moretti, Josephine T. Daub, Marc Robinson-Rechavi, and Laurent Keller. Patterns of positive selection in seven ant genomes. *Molecular Biology and Evolution*, 31(7):1661–1685, 2014.

[280] Thomas A. Oliver, David A. Garfield, Mollie K. Manier, Ralph Haygood, Gregory A. Wray, and Stephen R. Palumbi. Whole-genome positive selection and

habitat-driven evolution in a shallow and a deep-sea urchin. *Genome Biology and Evolution*, 2:800–814, 2010.

[281] Friso P. Palstra and Daniel E. Ruzzante. Genetic estimates of contemporary effective population size: what can they tell us about the importance of genetic stochasticity for wild population persistence? *Molecular Ecology*, 17(15):3428–3447, 2008.

[282] Friso P. Palstra and Dylan J. Fraser. Effective/census population size ratio estimation: a compendium and appraisal. *Ecology and Evolution*, 2(9):2357–2365, 2012.

[283] L. Chao and D. E. Carr. The molecular clock and the relationship between population-size and generation time. *Evolution*, 47(2):688–690, 1993.

[284] James S. Borges and Warren D. Johnson. Inhibition of multiplication of *Toxoplasma gondii* by human monocytes exposed to T-lymphocyte products. *The Journal of Experimental Medicine*, 141:483–496, 1975.

[285] W. A. Cresko, K. L. McGuigan, P. C. Phillips, and J. H. Postlethwait. Studies of threespine stickleback developmental evolution: progress and promise. *Genetica*, 129(1):105–26, 2007.

[286] J. A. Darling, A. R. Reitzel, P. M. Burton, M. E. Mazza, J. F. Ryan, J. C. Sullivan, and J. R. Finnerty. Rising starlet: the starlet sea anemone, *Nematostella vectensis*. *Bioessays*, 27(2):211–21, 2005.

[287] Matthew C. Fisher, Gina L. Koenig, Thomas J. White, Gioconda San-Blas, Ricardo Negroni, Isidro GutiÃ©rrez Alvarez, Bodo Wanke, and John W. Taylor. Biogeographic range expansion into South America by *Coccidioides immitis* mirrors New World patterns of human migration. *Proceedings of the National Academy of Sciences of the United States of America*, 98(8):4558–4562, 2001.

[288] D. C. Jeffares, T. Mourier, and D. Penny. The biology of intron gain and loss. *Trends in Genetics*, 22(1):16–22, 2006.

[289] Peter D. Keightley and Adam Eyre-Walker. Deleterious mutations and the evolution of sex. *Science*, 290(5490):331–333, 2000.

[290] Szilvia Kusza, Tomasz Podgorski, Massimo Scandura, Tomasz Borowik, Andras Javor, Vadim E. Sidorovich, Aleksei N. Bunevich, Mikhail Kolesnikov, and Bogumila Jedrzejewska. Contemporary genetic structure, phylogeography and past demographic processes of wild boar *Sus scrofa* population in central and eastern Europe. *PLoS ONE*, 9(3):e91401, 2014.

[291] P. Lemaire, W. C. Smith, and H. Nishida. Ascidians and the plasticity of the chordate developmental program. *Current Biology*, 18(14):R620–31, 2008.

[292] Nicholas D. Levsen, Peter Tiffin, and Matthew S. Olson. Pleistocene speciation in the genus *Populus* (*Salicaceae*). *Systematic Biology*, 61(3):401–412, 2012.

[293] E. D. Maloney and H. E. Kaufman. Multiplication and therapy of *Toxoplasma gondii* in tissue culture. *Journal of Bacteriology*, 88:319–321, 1964.

[294] A. Mira and N. A. Moran. Estimating population size and transmission bottlenecks in maternally transmitted endosymbiotic bacteria. *Microbial Ecology*, 44(2):137–143, 2002.

[295] R. E. Morales Vargas, P. Ya-Umphan, N. Phumala-Morales, N. Komalamisra, and J. P. Dujardin. Climate associated size and shape changes in *Aedes aegypti* (diptera: *Culicidae*) populations from Thailand. *Infection, Genetics and Evolution*, 10(4):580–585, 2010.

[296] S. Paland, J. K. Colbourne, and M. Lynch. Evolutionary history of contagious asexuality in *Daphnia pulex*. *Evolution*, 59(4):800–813, 2005.

[297] S. Subramanian. Nearly neutrality and the evolution of codon usage bias in eukaryotic genomes. *Genetics*, 178(4):2429–2432, 2008.

[298] D. S. Suman, S. N. Tikar, M. J. Mendki, D. Sukumaran, O. P. Agrawal, B. D. Parashar, and S. Prakash. Variations in life tables of geographically isolated strains of the mosquito *Culex quinquefasciatus*. *Medical and Veterinary Entomology*, 25(3):276–288, 2011.

[299] Sandrine Trouve, Pierre Sasal, Joseph Jourdane, Francois Renaud, and Serge Morand. The evolution of life-history traits in parasitic and free-living platyhelminthes: a new perspective. *Oecologia*, 115(3):370–378, 1998.

[300] L. Venturini, A. Ferrarini, S. Zenoni, G. B. Tornielli, M. Fasoli, S. Dal Santo, A. Minio, G. Buson, P. Tononi, E. D. Zago, G. Zamperin, D. Bellin, M. Pezzotti, and M. Delledonne. *De novo* transcriptome characterization of *Vitis vinifera* cv. *Corvina* unveils varietal diversity. *BMC Genomics*, 14:41, 2013.

[301] S. Gelfman, D. Burstein, O. Penn, A. Savchenko, M. Amit, S. Schwartz, T. Pupko, and G. Ast. Changes in exon-intron structure during vertebrate evolution affect the splicing pattern of exons. *Genome Research*, 22(1):35–50, 2012.

[302] Peter J. Shepard, Eun-A Choi, Anke Busch, and Klemens J. Hertel. Efficient internal exon recognition depends on near equal contributions from the 3' and 5' splice sites. *Nucleic Acids Research*, 39(20):8928–8937, 2011.

[303] D. A. Sterner, T. Carlo, and S. M. Berget. Architectural limits on split genes. *Proceedings of the National Academy of Sciences of the United States of America*, 93(26):15081–15085, 1996.

[304] Kristi L. Fox-Walsh, Yimeng Dou, Bianca J. Lam, She-pin Hung, Pierre F. Baldi, and Klemens J. Hertel. The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proceedings of the National Academy of Sciences of the United States of America*, 102(45):16176–16181, 2005.

[305] C. Fields. Information content of *Caenorhabditis elegans* splice site sequences varies with intron length. *Nucleic Acids Research*, 18(6):1509–1512, 1990.

[306] M. Weir and M. Rice. Ordered partitioning reveals extended splice-site consensus information. *Genome Research*, 14(1):67–78, 2004.

[307] D. M. Kupfer, S. D. Drabenstot, K. L. Buchanan, H. Lai, H. Zhu, D. W. Dyer, B. A. Roe, and J. W. Murphy. Introns and splicing elements of five diverse fungi. *Eukaryotic Cell*, 3(5):1088–1100, 2004.

[308] C. N. Dewey, I. B. Rogozin, and E. V. Koonin. Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns. *BMC Genomics*, 7:311, 2006.

[309] Feng-Chi Chen, Sheng-Shun Wang, Chuang-Jong Chen, Wen-Hsiung Li, and Trees-Juen Chuang. Alternatively and constitutively spliced exons are subject to different evolutionary forces. *Molecular Biology and Evolution*, 23(3):675–682, 2006.

[310] L. Collins, D. Penny, and Smbe Tri-National Young Investigators. Proceedings of the SMBE tri-national young investigators' workshop 2005. investigating the intron recognition mechanism in eukaryotes. *Molecular Biology and Evolution*, 23(5):901–10, 2006.

[311] S. H. Schwartz, J. Silva, D. Burstein, T. Pupko, E. Eyras, and G. Ast. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Research*, 18(1):88–103, 2008.

[312] Jessica A. Thomas, John J. Welch, Robert Lanfear, and Lindell Bromham. A generation time effect on the rate of molecular evolution in invertebrates. *Molecular Biology and Evolution*, 27(5):1173–1180, 2010.

[313] N. P. Kandul and M. A. Noor. Large introns in relation to alternative splicing and gene evolution: a case study of *Drosophila bruno-3*. *BMC Genetics*, 10:67, 2009.

[314] G. Ast. How did alternative splicing evolve? *Nature Reviews Genetics*, 5(10):773–782, 2004.